DECLARATION - USA PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name;

I believe I am an original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled INFRACTION FINGERPRINT ANNOTATIONS FROM PROTEIN STRUCTURE MODELS; the specification of which was filed on **August 20, 2001** as Application Serial No. **09/933,580**.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above;

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, § 1.56;

I hereby claim the benefit under Title 35, United States Codes § 119(e) of any United States provisional application(s) listed below.

Application No.: 60/226,327 Filing Date: August 18, 2000

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful, false statements may jeopardize the validity of the application or any patent issued thereon.

Full name of first inventor: **Sandor Szalma**

Inventor's signature _____

Date 11/30/01

Residence: **12987 Caminito Bautizo, San Diego, CA 92130**

Citizenship: HUNGARIAN

Post Office Address: **12987 Caminito Bautizo, San Diego, CA 92130**

Full name of second inventor: **Mariusz Milik**

Inventor's signature Mariusz Milik

Date 12.03.2001

Residence: **3511 Stetson Ave, San Diego, CA 92122**

Citizenship: POLISH

Post Office Address: **3511 Stetson Ave, San Diego, CA 92122**

Full name of third inventor: **Krzysztof Olszewski**

Inventor's signature

Date 3 Dec. 2001

Residence: **12260 Porcelina Court, San Diego, CA 92131**

Citizenship: POLISH

Post Office Address: **12260 Porcelina Court, San Diego, CA 92131**

Full name of fourth inventor: **Lisa Yan**

Inventor's signature

Date Dec 3, 2001

Residence: **11960 Black Mountain Road, #52, San Diego, CA 92129**

Citizenship: U.S.A.

Post Office Address: **11960 Black Mountain Road, #52, San Diego, CA 92129**

Full name of fifth inventor: **Azat Badretdinov**

Inventor's signature _____

Date 11/30/1

Residence: **11584 Windcrest Lane, San Diego, CA 92129**

Citizenship: RUSSIAN FEDERATION

Post Office Address: **11584 Windcrest Lane, San Diego, CA 92129**

Full name of sixth inventor: **Scott Kahn**

Inventor's signature _____

Date 12-11-01

Residence: **14259 Woodcreek Road, Poway, CA 92064**

Citizenship: USA

Post Office Address: **14259 Woodcreek Road, Poway, CA 92064**

Send Correspondence To:
KNOBBE, MARTENS, OLSON & BEAR, LLP
Customer No. 20,995

R:\DOCS\KEP\KEP-1248.DOC:lw
112101

Guide for Features Annotations

Part 3 - Domains, Regions and Products

Version 8
28-SEP-1998

"Domain" Record

The format for the "Domain:" record is

> *"Domain:" ["(or "hyphenated pairs")"] domain name ["("form")"] ["#status" status] "<" tag ">"*

This record should be generally be applied to a single hyphenated pair. A "domain" carries the connotation of having some degree of spatial coherence, that is, secondary or tertiary structure. Separate segments of sequence that together form the same domain should be placed in the same record. Separate segments of sequence that form spatially distinct domains that happen to have the same description should be placed in separate records.

We have attempted to standardize most "Domain" records, but this format is still somewhat variable. Here we set forth some very general guidelines pertaining to certain types of domains.

Back to Top

General Principles

Use the same name for the same kind of domain. Insofar as possible, use the same or similar tags for the same kind of domain. Domain names should be INFORMATIVE; avoid names such as "first", "A", "II", etc. A domain or region should be annotated only when it is biologically significant and the name should reflect that interesting structural or functional property. Names that are obvious or used only for the

convenience of particular authors should be suspect.
Do not include enumeration within the names given to repeated domains of the same type within the same sequence. This results in needless proliferation of names that are all the same except for a number or letter. The enumeration should be in the tag instead.

The boundaries of domains are assumed to have "predicted" status and are understood to be not necessarily precise; usually no additional indication of uncertainty is needed. If there is considerable uncertainty, for example if any of three Mets might be the initiator, this is indicated by an initial parenthetical phrase. For example,

Domain: (or 5-32 or 11-32)

The "or" form should be avoided whenever possible.

The boundaries of homology domains are understood to be more or less arbitary and defined on the basis of sequence similarities; do not put a status on such domains.

The question of what types of regions we should call domains is still under discussion.
Each instance of a given kind of domain within a sequence should have a separate domain record, thus use

20-42/Domain: transmembrane #status predicted
50-72/Domain: transmembrane #status predicted

and do not use

20-42,50-72/Domain: transmembrane #status predicted

In some cases a single 3-dimensionally defined domain does consist of separated segments of sequence, and a list of ranges may appear in such cases but this is rare.

Back to Top

Signal sequences and transit peptides

These domains have been standardized in PIR. Please follow the format given below for the simple cases; in more complex cases, use the examples as a guide. A form must appear with a transit peptide. Tags are required, but these examples are suggestions.

"Domain: " ["(or"hyphenated pair ["or" hyphenated pair ...]")"] "signal sequence" ["(fragment)"] ["#status" status] "<SIG>"

"Domain:" ["(or" hyphenated pair ["or" hyphenated pair ...] ")"] "transit peptide ("form")" ["(fragment)"] ["#status" status] "<TNP>"

form is "mitochondrion" | "chloroplast" | "amyloplast" | "chromoplast" | "cyanelle" | "glyoxysome" | "hydrogenosome" | "plastid" | "thylakoid"

Examples:

Domain: signal sequence #status predicted <SIG>
Domain: signal sequence (fragment) #status experimental <SIG>
Domain: transit peptide (amyloplast) #status predicted <TNP>
Domain: transit peptide (chloroplast) #status predicted <TNP>
Domain: transit peptide (chloroplast) (fragment) #status experimental <TNP>
Domain: transit peptide (mitochondrion) #status predicted <TNP>

The "or" form should be avoided whenever possible.

Domain: (or 1-15) signal sequence (fragment) #status predicted <SIG>
Domain: (or 1-43 or 1-49) signal sequence #status predicted <SIG>

When the boundary between a signal sequence and the following domain has not been determined or predicted, use a record like one of these:

Domain: signal sequence and propeptide #status predicted <SIG>
Domain: signal sequence (fragment) and propeptide #status predicted <SIG>

When more than one protein product is presented in an entry, use this format.

"Domain: signal sequence (of "product name") ("status") <SIG>"

For example

Domain: signal sequence (of membrane glycoprotein E1) #status predicted <SIG>

Back to Top

Membrane-crossing regions

These are currently annotated as domains.

Domain: transmembrane #status predicted <TMM>
Domain: transmembrane beta strand #status predicted <TMM>
Domain: transmembrane helix #status experimental <TMM>

Try to be consistent in assigning boundaries of transmembrane domains within a group of closely replated proteins. Lacking any other criteria, use the minimum range suggested by the ALOM program. The preferred tags are "<TMM>" when there is only one, and "<TM1>", "<TM2>", etc., when there are more than one. When there are more than nine, use tags like "<TM01>".

[BLACK] Do not use the following kinds of names for transmembrane domains:

"transmembrane 2" or "transmembrane II" (the numbers should not be part of the name)
"transmembrane domain"

"transmembrane region"
"membrane-spanning segment"
"potential transmembrane sequence"
"membrane anchor domain"

[GRAY]
The following cases also appear and are under review.

Domain: intramembrane
Domain: membrane anchor
Domain: membrane associated
Domain: membrane insertion
Domain: membrane-bound
Domain: transmembrane amphipathic helix #status predicted

Back to Top

Homology domains

Homology domains form a special class. They are distinguished by the property that those of a given type (with the same name) are homeomorphic and share sequence homology although they are found in different (nonhomeomorphic) proteins. The names of such domains end with the word "homology". Many such domains are homologous with most of the entire length of some other protein, in which case they may be named after such a protein, either exactly ("trypsin homology") or with a more general designator ("protein kinase homology"). Other domains have, so far, been found as domains within multidomain proteins ("homeobox homology"). Some are named to indicate that they are repeated in a certain protein ("complement factor H repeat homology"). The conversion of homology domain names to include the terms "homology" or "repeat homology" is still incomplete.

The boundaries of homology domains should be consistent with an alignment of representative domains of the named type. Dr. Barker is collecting such alignments. A preferred tag will be assigned for each type of homology domain. Some examples:

Domain: basic proteinase inhibitor homology <BPI>
Domain: cytochrome b5 core homology <CB5>
Domain: protein kinase homology <KIN>
Domain: calmodulin repeat homology <EF1>

Note that some domains with names of proteins may have been assigned not by sequence homology but by predicted activity. Defining a homology domain is preferable to a name that predicts structure ("EF hand") or function ("calcium-binding") because structure may be distorted or function lost in homologous domains. Do NOT add the word "homology" without affirming that there is sequence homology! Please read carefully the discussion and proposal of the use of "Domain" and "Region" records for repeated sequence elements.

[BLACK] Do not use a status for this type of domain. They are only assigned by homology, which is always an inference of predicted status and never experimental. Boundaries are understood to be somewhat a matter of human judgement.

[BLACK] The following names are NOT acceptable:

"alpha chain homolog"
"complement binding protein-related"
"endozepine-like"
"homology with Ig C region domains"
"malK protein homolog 1"
"lipoyl domain 1 #status predicted"

In this last example, it's not clear whether it is an homology domain or a function prediction. The word "domain" is always superfluous. Domains should not be enumerated.

To denote regions that are under consideration as homology domains, it has become acceptable practice to annotate them as a "similarity" like

Domain: platelet-derived growth factor chain B similarity

with the understanding that they will be changed at a later date.

Back to Top

Repeat Domains

Domains that are repeated in a protein should be names as homology domains if they are also known to occur in diverse proteins. Otherwise, they may be named for the specific protein or an example of it omitting the term "homology", for example "CDC23 repeat".

Back to Top

Miscellaneous Rules

Use a hyphen before "binding" as in "cAMP-binding" when it is used attributively, that is as an adjective with a following noun. For example,

Domain: DNA-binding core #status predicted

Domain: alpha-actinin actin-binding domain homology

Otherwise, if it is used nominatively, as a name with no following noun, do not use a hypen. For example,

Domain: DNA binding #status predicted Region: actin binding #status predicted

If required, use "fragment". It appears before the status (if used) and tag.

Always try to use names that are at least 3 characters in length.

[BLACK]Old "Duplication" records cannot be used. Any features of this type should be entered as "Region" or "Domain" records in accordance with the discussion below.

Following these guidelines will at least reduce the heterogeneity in the current database and make it more easy to convert.

Duplications are of two major types: short repeats (usually tandem) and longer domains. There is no firm cut-off without studying the situation; as a guideline, we could try 25 or fewer residues is a repeat, 50 or more is a domain, and in between it must be a domain if such domains exists in other types of proteins (e.g., EGF-like) but may be treated as a tandem repeat if it is unique to this type of protein.

Back to Top

Repeats

Repeats are very to fairly short, usually occur in tandem, and the pattern is often, but not always, specific to this type of protein. The annotation to use is *"Region:" record* Use a hyphenated pair for the entire region in the location field, "22-300", and do not give the boundaries of individual repeats. Several unconnected regions may be listed if they contain the same pattern, "22-100, 200-298" (note: we don't have the permission to use a semi-colon yet; hopefully it comes very soon!) If there is a list, then all the other information within the record must be applicable to the entire list.
In the description field, use the following format

n "-residue repeats" ["("sequence pattern")] [", " descriptive phrase]

For example

Region: 11-residue repeats (D-P-A-K-A-S-Q-G-G-L-E)

"n" is the typical number of residues in the repeat pattern and the number of repeats is not given.
A sequence pattern is a simple representation of the canonical pattern using the single-letter code separated by hyphens and when necessary alternatives are indicated for only the most common residues separated by a slash. For example, *(A-C-D/E-F-G)* No tag is usually used with the "Region" record. "tandem repeat" is used as a KEYWORD if the repeats are tandem. "repeat" may be allowed as a KEYWORD for non-tandem repeats.

Back to Top

Domains

It is suggested that the longer domains should be annotated as domains, individually represented, and tagged so that these subsequences can be retrieved. All domains of the same type should be given exactly the same name. For domains defined by homology, there will be eventually an alignment of

selected examples which, in effect, is the definition of the domain. Dr. Barker is curator for homology domains and welcomes any such alignments from other PIR-International staff for the definition of new domains or for the standardization of currently heterogeneous domains.

Back to Top

"Product" Records

A Product is any relatively stable (i.e. isolatable) peptide chain, including chains that experience cleavage of a precursor form and remain bound together in the same molecule. This definition has several implications.

Some sequence elements previously identified as "Peptide" are probably not stable and will not fit the proposed definition of "Product". You may use "Domain" or "Region" for these. Activation peptide are normally annotated as a "Domain" unless they have been isolated and appear to be physiologically significant. What can usually be easily determined is what segments are present in the final mature protein(s) and what segments are removed.

[BLACK] Do not use Product records like

20-50,70-90/Product: mcguffin A and B chains #status experimental <MAT>

Several options are available, and there are good examples where it has been necessary to use one or the other of these forms. You may represent the chains in two separate "Product" features.

20-50/Product: mcguffin chain A #status experimental <ACH>

70-90/Product: mcguffin chain B #status experimental <BCH>

It is also possible to present a single "Product" feature and two "Domain" features, especially when the chains are covalently linked and only a single molecular entity with one molecular weight actually exists.

20-50,70-90/Product: mcguffin #status experimental <MAT>

20-50/Domain: mcguffin chain A #status experimental <ACH>

70-90/Domain: mcguffin chain B #status experimental <BCH>

[The use of the second approach is evident in annotating protein splicing.] Do not mix these forms in the same entry, and try to standardize them across a family.

So far there has been little standardization of "Product" records; however, the following guidelines should be used.
Do not use "Product" for a segment that has insufficient lifetime to be isolated.

A name in a "Product" feature should repeat the protein name as given in the entry title or a name in the "Contains" record, usually omitting "precursor" and including a chain designation. Version and clone designations may be omitted. This may be enforced at a later date.

Chain designations that are words or Greek letters should precede the word "chain" and designations that are English letters, numbers (Arabic or Roman) or combinations them should follow the word "chain": thus,

"chain B2"
"chain IV"
"pi chain"
"heavy chain"
"catalytic chain"

The tag is required. Use "<MAT>" for a single mature product.

If you can determine that at least both boundaries of a product have been experimentally determined AS PROTEIN with substantially enough of the portion between to leave little doubt that additional processing or splice forms do not occur, then use the status "#status experimental". Use "#status predicted" if the boundaries are assigned by homology or the sequence is determined substantially as nucleic acid. The experimental determination of only one end (almost always the amino end) is not sufficient to justify use of "#status experimental" for an entire "Product" feature because protein splicing, alternate transcripts, frame-shift errors and carboxyl-terminal propeptide processing introduce too many uncertainties.

[BLACK] Do not use "amino end of" and "carboxyl end of". Instead use the by modifier "(fragment)".

Back to Top

"Region" Record

This record remains generally unstandardized at this time to allow the annotation of new features that are not yet well-understood or standardized. A "Region" should probably carry the only the connotation of being contiguous sequence, as opposed to the spatial connotation of a "Domain". The following guidelines should be followed:

The tag is not usually used.

Status is often not appropriate.

See the discussion elsewhere for how to handle regions of tandem repeats.

The word "rich" should be appended with a hyphen. The word "binding" should be appended with a hypen if it is used as an adjective, and it should not have a hyphen if it is not followed by a noun.
Do not use the word "region" in the description; no "Region: xxx region".

[BLACK] Avoid using expressions which match other record types, such as:

Region: active site

Region: extracellular domain

This first expression should especially not be used if specific residues are listed. It should either be annotated as an "Active site" or as

Region: catalytic

The second would be better as Domain: extracellular #status predicted
Regions of a specific type of secondary structure should not be annotated. In the NRL_3D database only, the PDB HELIX, TURN and SHEET features are converted to PIR "Region" features. The definitions and descriptions will use the PDB annotations in parentheses.

Region: helix (right hand alpha)

Region: turn (type II)

Region: beta sheet

Region: beta barrel

No other PIR databases should have entries with such conformational information annotated. The feature

Domain: beta barrel

is acceptable.

Motifs or patterns combining various types of secondary structure may be annotated as "Regions". For example, *Region: helix-turn-helix motif <HTH>*

Do not use the word "motif" except for defined or accepted sequence motifs. Use the word "pattern" instead.

Do not use "#status predicted" for any feature that is defined by a sequence motif or pattern. Do not use something like

Region: pentapeptide motif (X-F-X-F-G) #status predicted

This is nonsensical because either the pattern is in the sequence or it is not. Instead use

Region: pentapeptide motif (X-F-X-F-G)

Unfortunately it becomes more difficult to appreciate this rule when the name given to the motif is supposedly descriptive of a function. In cases like

Region: DNA-binding motif (K/R-G-R-G-R-P)

it is very tempting to use "#status predicted". But does the status mean that the property of DNA binding

is predicted, or only that a motif is present? If the motif is present, it certainly isn't predicted, it is experimentally observed. But putting "experimental" would suggest that "DNA-binding" is not just a name but an observation. Don't be confused or confusing; never use a status with "motif", "pattern", "homology", "similarity", etc.

Back to Top

Suggestions for Annotators

Annotators may wish to use this checklist in preparing an annotation.
Usually the annotation should be the same as an annotation already in the database. Check for the feature in other database entries. Only these record types should be used:

Active site:
Binding site:
Cleavage site:
Cross-link:
Disulfide bonds:
Domain:
Inhibitory site:
Modified site:
Product:
Region:

The use of only these types is enforced in PIR databases.

Except for the special cases of "selenocysteine" and "N-formylmethionine", standard 3-letter residue codes should appear after the colon of "Active site", "Cleavage site" and "Inhibitory site" records, and in parentheses immediately after the first name in "Binding site", "Cross-link" and "Modified site" records. Be certain the residue code appears, that the residue has the correct number and that it corresponds to the proper residue in the sequence. This identity check is enforced in PIR databases.
Check that all other required fields are present and in the preferred order. The status should always be added to new entries in these records:

Active site:
Binding site:
Cleavage site:
Cross-link:
Disulfide bonds:
Inhibitory site:
Modified site:
Product:

A status may be appropriately used in only some "Region" and "Domain" records. If the extent field is used, only the word "partial" should appear, it should be placed immediately before the status and the status should be "experimental", not "predicted"

Check your spelling and punctuation. Spelling errors in chemical terms can be especially difficult to

catch. When appropriate, check that the names in "Product" records correspond to names in the title or "Contains" record.

Check that there are unique tags on all "Product" and "Domain" records, and that they are different from other tags in the entry. Tags are not required on any other types of features.

Back to Top

 Back to Features Table of Contents.

E-Mail comments or suggestions about this site.

Revised 11/12/98

Position-specific annotation of protein function based on multiple homologs

Miguel A. Andrade
European Molecular Biology Laboratory,
69012 Heidelberg, Germany
andrade@embl-heidelberg.de

Abstract

In this work I present an algorithm for deriving position-specific protein functional annotations. The input is based on the results of a sequence similarity search of a query sequence against a sequence database. Strings of words are extracted from the descriptions of the proteins, and the correlation between proteins having the same descriptors and amino acid conservation is used to compute a score that indicates which descriptor is likely to best describe the function of each particular residue. Analysis of the score curves and comparison of different functions allows an easy detection of parts of the sequence associated with different functions. Different levels of functional specificity can be compared, allowing the choice of the one that best suits the function of the protein. Immediate applications of this algorithm are, support for (automated) methods of protein functional annotation, and database coherency checking.

Introduction

The advent of genome projects is producing increasing numbers of putative new proteins. The biochemical functional characterization of all of these proteins is an impossible task. However, Bioinformatics uses the combination of algorithms and knowledge on proteins to allow a preliminary computer-based functional characterization, which is much faster and less expensive (Andrade & Sander 1997). These algorithms are normally based on the principle that sequence similarity between proteins corresponds to some functional similarity. Function from characterized proteins is therefore transferred to proteins to be characterized. This transference is affected by a series of problems related to the limitations of the methods used and the intrinsic complexities of protein function (Bork & Bairoch 1996; Bork & Koonin 1998; Galperin & Koonin 1998; Andrade *et al.* 1999).

Failures of methods or careless interpretation of their results can generate incorrect functional assignments. Even worse, these errors can be introduced as 'truth' in databases, generating new errors, as other proteins may be characterized from those erroneously assigned. This problem aggravates when more or less automated methods are used for the process of annotation. However, a certain degree of automation is desirable when confronted with large numbers of proteins to be characterized. Curation of existing database data would also benefit from better automated methods for annotation.

As discussed elsewhere (Andrade *et al.* 1999), there are three main problems in functional transfer: (i) wrong annotations — the protein used for transference has incorrect functional annotation; (ii) false positives — the sequence similarity used for the transference is too weak and does not corresponds to real functional similarity; (iii) inaccurate transfers that can be associated to (iiia) the domain problem — the sequence similarity corresponds to regions of the protein that are not involved in the transferred functionality, or to (iiib) functional hierarchy — the sequence similarity is not strong enough to account for the function transferred, although there is functional similarity between the proteins at a lower specificity level.

There are two main sources for these errors: (i) the transfer is done from only one protein without considering the information from other members of the family, and (ii) the transfer is done from protein to protein as a whole, without taking into account protein fragments (or domains).

In this work, I suggest a representation relating sequence similarity to functional similarity. In this particular application, sequence similarity is taken from sequence to sequence similarity searches [specifically from the gapped BLAST program (Altschul *et al.* 1997)] and function is taken from the protein descriptions (unrelated to specific positions on the protein) as given in the SWISSPROT (Bairoch & Apweiler 1999) or SPTREMBL databases 'DE' field. The algorithm is general enough to allow the use of other inputs. However, protein descriptions are today the only general description of function available for most of the characterized sequences, even if inexact, unformatted and heterogeneous. The system is expected to be able to cope with a low level of noise and errors.

Gapped BLAST has been preferred to profile iterated BLAST [PSI-BLAST (Altschul *et al.* 1997)] because

the latter is more appropriate for iterative searches likely to focus on fractions of the set of homologs. In this case, it is enough to get an overview of the sequence space around the query without exploring the remote homologs with more sensitive searches.

The information on residue conservation in sequences having a given functional descriptor is used for scoring the likelihood of this protein function to describing each residue of the query sequence. The scores compose a mathematical function along the sequence that is scanned for regions of the sequence corresponding to values above a given threshold. The set of functional assignments and their locations provide a rich and simple functional overview of the query protein which can be used for identifying the appropriate functional transfer.

Method

Sequence similarity

Given a query sequence, a BLAST sequence similarity search against a protein database gives a list of hits of this to a series of proteins (above a certain very low, normally non-significant, cut-off of similarity). The BLAST result is pre-processed with MView (Brown, Leroy, & Sander 1998). This tool converts the results of a sequence database search into a multiple alignment of hits stacked against the query. The resulting sequence is a construct and therefore may differ slightly from the real hit sequence. MView is also used to remove highly redundant sequences (with more than 95% sequence identity): the algorithm relies on the differences in conservation between regions implicated in different functions, and these differences cannot be appreciated unless a certain degree of divergence exists between the sequences to compare.

Analysis of functional information

Analysis of the functional information in the description lines attached to each sequence is carried out in 5 steps (a-e):

a) clean annotations. Any capitals are translated into small letters. Symbols are translated into spaces.

b) eliminate low frequency words. Since I am going to analyse sequence conservation patterns in sequences having similar annotations, words scarcely used in the whole set of descriptions have to be dismissed as they would not provide significant results. In this application, I used words present more than 5 times. This enormously simplifies the following steps.

c) find word units. Find all possible strings of consecutive words (word units) that are associated with at least a minimum number of proteins (again in this case, more than five proteins). The distribution (presence/absence) of all word units in the set of proteins is annotated.

d) word unit simplification. Word units contained in other word units and having exactly the same distribution are eliminated as redundant. Overlapping word units having the same distribution are joined.

e) elimination of word units composed of only numbers or single letters. They have no meaning. For example, the string "ec 2 7 1 4" has a meaning (hexokinase) and the sub-strings "ec 2" or "ec 2 7 1" have also a meaning (transferase and phosphotransferase with an alcohol group as acceptor, respectively). On the contrary, "7 1" without the "ec" has no meaning. Similar considerations apply to single characters coming from chemical formulae or to those describing functional specificity.

The final result of this stage is the word unit usage matrix. The columns of this matrix represent word units and the rows proteins. Entries indicate whether a protein contains a given word unit in its description or not.

Scoring function

I am going to describe the scoring function used for the evaluation of the likelihood of a word-unit assignment to a given residue of the query sequence. This score depends on three parameters: a measure of the degree of disorder in the amino acid distribution of the hits at the given position – which I will denote entropy although it is not the physical magnitude – (s), the fraction of proteins having the word unit that hit the query sequence at the position (g) and the fraction of proteins having the word unit that match the residue of the query sequence at the position (f). Note that, BLAST gives fragments from homolog sequences that hit the query; not all the amino acids in these hit-regions match exactly the query.

The amino acid distribution at a given position of the MView alignment (corresponding to a position in the query sequence) is the set of counts of different amino acids present in a column (X_i for $i = 1, \ldots, 20$) of the set of proteins having a given word unit in their descriptions. The entropy of this distribution is computed in the following way:

$$s = 1 - \frac{\sum_{i=1}^{20} X_i(X_i - 1)}{n(n-1)} \qquad (1)$$

where n is $\sum_{i=1}^{20} X_i$. Note that $0 \leq s \leq 1$ (0 for a complete conserved position, and 1 for a completely variable position).

Dreaming of a scoring function

I will introduce some intuitive rules that the scoring function should fulfill (see Fig. 1). See the discussion for another proposed modes of defining a score.

Rule a) given two positions with the same entropy, a higher matching fraction is preferable. **Rule b)** for a high entropy, the fraction matched should not matter. **Rule c)** if the fraction matched is reasonably high, a position with lower entropy should score better. **Rule d)** positions with very low entropy have to be matched (otherwise they score negatively). **Rule e)** if the fraction of proteins with the word unit that hit the protein is very low, the score should be less significant.

The function I have chosen is (see inset in Fig. 1):

$$\epsilon(s, f, g) = g(1 - s)(a + bf^c) \quad (2)$$

where ϵ may take either positive or negative values, for good and bad positions, respectively. In the following I am going to show how Eq. 2 fulfills the proposed rules and what is the meaning of the a, b and c constants.

The entropic component $(1 - s)$ makes the function approach zero when the entropy approaches 1 (rule b). A similar effect is played by the g term (rule e). Provided that $b > 0$, the scoring function ϵ increases with f (rule a).

The significance of the a and b constants can be understood in the situation where the position to match is totally conserved (i.e., $s = 0$ and $g = 1$). If the position is matched by the query sequence ($f = 1$) then $\epsilon = a+b$ (see Eq. 2). If the position is not matched by the query ($f = 0$) then $\epsilon = a$.

If $a < 0$ and $a + b > 0$ there is a crossing point (γ) of the scoring function ($\epsilon = 0$) independent on the values of g and s (see inset in Fig. 1). This can be chosen to set the matched fraction f that reverses the effect of the entropy. At $f < \gamma$, lower entropy values give lower scores (rule d). At $f = \gamma$ the score is zero (irrespective of the entropy value). At $f > \gamma$, lower entropy values give higher scores (rule c).

The c exponent value dependence on the other parameters can be easily computed as:

$$c = \frac{\ln\left(-\frac{a}{b}\right)}{\ln \gamma} \quad (3)$$

In this application I have set $\gamma = 0.1$ (a query residue matching 10% of the residues of the amino acid distribution scores zero), $a = -0.5$ (not matching a completely conserved position scores -0.5), and $b = 1.5$ (matching a completely conserved position scores $b - a = 1$).

Graphical representation and functional region detection

The resulting scoring functions for each word unit are smoothed along the sequence using a Gaussian filter (with a width of $\sigma = 20$ residues).

In order to make a preliminary annotation, regions of the sequence with good transference probability and a minimum length (50 residues in this case) are reported. Three score thresholds are defined: clear, tentative, and marginal [following GeneQuiz reliability nomenclature (Scharf *et al.* 1994)] (here 0.03, 0.01 and 0.005, respectively). These regions are reported by word unit, start and stop in query sequence, reliability class of the putative functional assignment and mean score.

Examples

I have used some tricky cases [some constituting GeneQuiz failures and discussed before (Andrade *et al.* 1999)] to illustrate the possible solutions to the transference problems described in the introduction. Table I describes the functions annotated for these sequences.

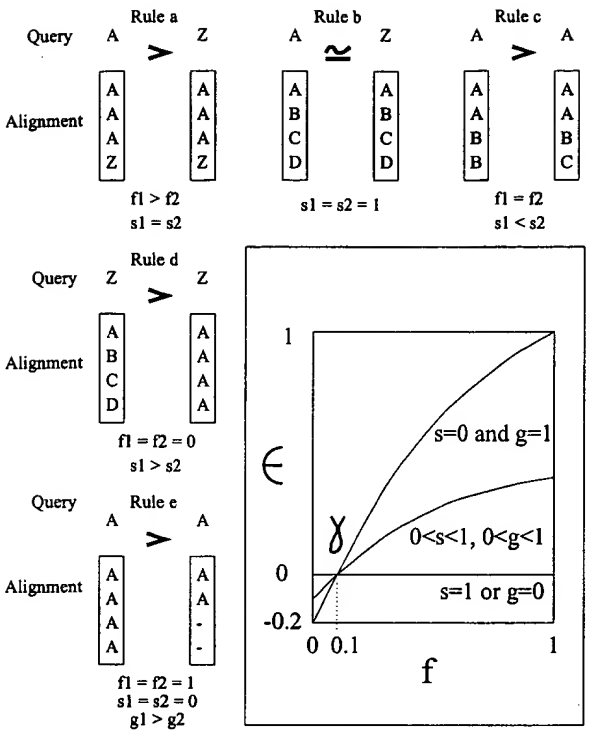


Figure 1: Graphic illustration of the five rules that the scoring function ϵ (Eq. 2) should fulfill (see text). The function is depicted in the inset (bottom right). For each rule, two examples of a matching of a single amino acid of the query (top) against a column of the alignment (vertical boxes) are presented, with a 'should be' score comparison in the middle and some considerations on the respective f, g, and s values (bottom). For example, the interpretation of the "rule a" depicts an 'A' amino acid matching a 'AAAZ' set of amino acids which should score better than a 'Z' amino acid matching the same 'AAAZ' set of amino acids.

Wrong annotation

`SW:KMHC_DICDI`[1] is incorrectly annotated in the SWISSPROT database as "myosin heavy chain kinase" (all closely related homologs are diacylglycerol kinases, it has no homology to any myosin heavy chain kinase, and the entry itself displays the presence in the sequence of diacylglycerol kinase patterns absent in myosin heavy chain kinases).

This error is easily detected by the analysis (see Table I and Fig. 2a). The obvious annotation is diacylglycerol kinase. "myosin heavy chain kinase" does not even appear as a valid word unit. Note that the C-terminal region of the protein is left unannotated.

[1]The notation used for database entries is database:identifier with SW for SWISSPROT and SP for SPTREMBL.

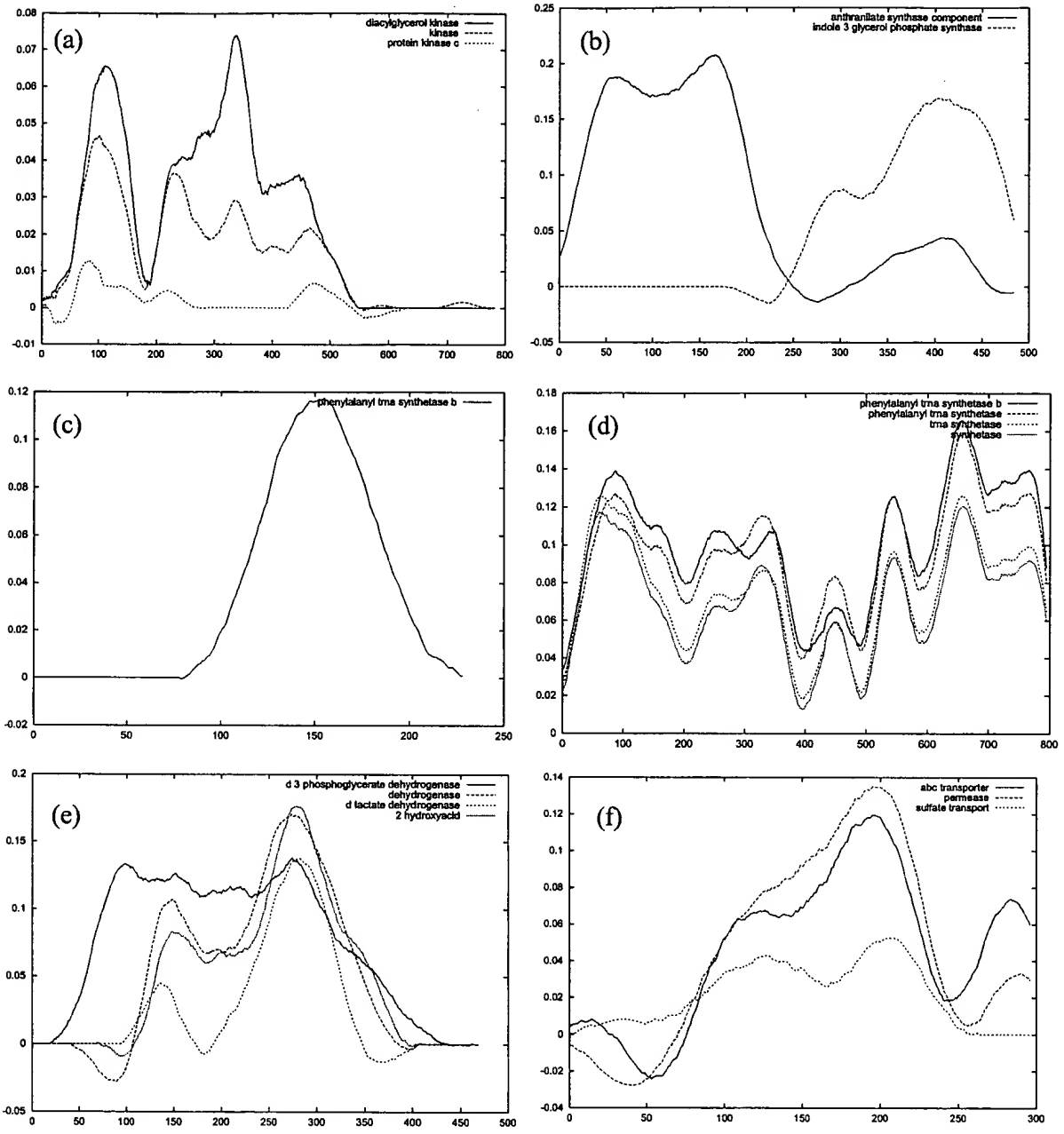


Figure 2: Plots of the score for functional transference for some of the examples depicted in Table I. Horizontal axis: query sequence position in residues. Vertical axis: functional transference score, ϵ. (a) SW:KMHC_DICDI. (b) SW:TRPG_YEAST (c) SW:Y449_MYCGE . (d) SW:SYFB_ECOLI. (e) SW:SERX_YEAST. (f) SP:O31520.

from	to	r.	score	word unit
			SW:KMHC_DICDI	
68	156	clear	0.050	diacylglycerol
215	380	clear	0.046	"
392	465	clear	0.032	"
66	158	clear	0.053	" kinase
213	465	clear	0.043	" "
68	142	clear	0.040	kinase
193	515	tent	0.021	"
61	122	clear	0.039	ec
203	364	clear	0.040	"
191	288	tent	0.023	" 2
415	478	tent	0.015	" "
			SW:TRPG_YEAST	
2	245	clear	0.144	anthranilate
356	445	clear	0.040	"
2	247	clear	0.146	" synthase
354	446	clear	0.041	" "
2	250	clear	0.147	" " component
348	449	clear	0.044	" " "
2	240	clear	0.131	synthase
2	214	clear	0.083	para aminobenzoate synthase
4	143	clear	0.065	" " " gl
155	206	clear	0.037	" " " "
246	483	clear	0.118	indole 3 glycerol phosphate synthase
29	104	tent	0.020	phosphate
253	483	clear	0.089	"
34	104	clear	0.052	carbamoyl phosphate
35	107	clear	0.057	" " synthase
62	204	clear	0.042	gmp
65	204	clear	0.040	" synthase
67	134	clear	0.040	" " glutamine
67	135	clear	0.040	" " " hydrolase
138	195	clear	0.037	" " " "
			SW:Y449_MYCGE	
100	201	clear	0.0814	phenylalanyl trna synthetase
104	202	clear	0.0828	" " " b
			SW:SYFB_ECOLI	
2	794	clear	0.107	phenylalanyl trna synthetase
2	794	clear	0.111	" " " b
2	794	clear	0.086	trna
2	794	clear	0.091	" synthetase
2	794	clear	0.087	synthetase
			SW:SERX_YEAST	
37	468	clear	0.174	d 3 phosphoglycerate
45	409	clear	0.101	" " " dehydrogenase
101	374	clear	0.102	dehydrogenase
107	363	clear	0.097	" ec
105	365	clear	0.096	" " 1
109	359	clear	0.108	ec
112	376	clear	0.096	2 hydroxyacid
203	337	clear	0.085	d lactate dehydrogenase
211	339	clear	0.094	" " " ec
113	362	clear	0.057	protein
108	171	clear	0.059	formate
223	295	clear	0.047	"
177	327	clear	0.045	hypothetical
			SP:034978	
2	431	clear	0.078	hypothetical
2	431	clear	0.072	protein
2	305	clear	0.065	kd protein i
347	430	clear	0.067	" " "

from	to	reliab.	score	word unit
			SP:031520	
13	295	clear	0.078	hypothetical
6	295	clear	0.076	" abc
6	295	clear	0.067	" " transporter
14	295	clear	0.088	abc
73	295	clear	0.082	" transporter
73	295	clear	0.078	probable
83	295	clear	0.072	" abc transporter
95	239	clear	0.065	" " " permease
75	295	clear	0.086	permease
33	254	clear	0.066	transport
31	251	clear	0.066	" system
88	235	clear	0.068	" " permease
145	257	clear	0.106	protein
85	158	clear	0.040	sulfate transport
174	228	clear	0.045	" "

Table I: Analysis of seven examples (see text). Valid word units and blocks of function detected from those with clear and tentative reliability. Graphs for all examples (except SP:034978) can be seen in Fig. 2.

Domain problem

SW:TRPG_YEAST is an example of multi-functional enzyme. It contains two proteins fused in one, which appear separately in many other organisms. The BLAST output clearly reflects this fact since two different blocks of hits associated to the different functionality appear clearly segregated. The analysis displays clearly the two separated functions (Table I and Fig. 2b). The crossing at about position 250 indicates the domain border.

SW:Y449_MYCGE illustrates a more complex example of domain problem. It is a small hypothetical protein (228 amino acids) which shares a C-terminal domain with many other proteins of various functions (Koonin *et al.* 1997). The closest hits are to those tRNA synthetases containing this domain. Transfer of this function to the query leaves the N-terminal 100 amino acids without annotation (see Table I and Fig. 2c).

The complementary analysis of one of close homologs would be necessary in this case to show that the domain is not specific for the transferred function. For example, the analysis of the closest homolog (SW:SYFB_ECOLI annotated as phenylalanyl-tRNA synthetase β chain) seems to indicate that the whole of the protein is required for this function (see Table I, Fig. 2d) and not only the region matching SW:Y449_MYCGE (N-terminal 200 amino acids). This case shows a nice correlation between increasing scores and increasing functional specificity.

Functional hierarchy

SW:SERX_YEAST is annotated as putative D-3-phosphoglycerate dehydrogenase. The analysis (see Table I,Fig. 2e) validates this annotation and shows, as in the previous case, the different levels of functional hierarchy. Note how annotations such as "2 hydroxyacid" and "dehydrogenase" have very similar plots. In this case, this is an indication of complementarity of the annotations: "2 hydroxyacid dehydrogenase" is

a generic definition that includes several homologs to the query such as the "D-3-phosphoglycerate dehydrogenase" and the "lactate dehydrogenase".

A more complicated case of functional hierarchy is shown in the analysis of SP:031520. A previous GeneQuiz analysis reported a too specific function for this protein (Andrade *et al.* 1999): "lactose permease". In this analysis, descriptions such as "lactose" or even the more general "sugar permease" are not even recorded as valid word units given the low incidence of the word in the descriptions of the homologs. The analysis (see Table I and Fig. 2f) gives a number less specific general descriptors such as "abc transporter" or "permease". More specific descriptions such as "sulfate permease" score much lower. Note that the 100 N-terminal amino acids of the query sequence remain unannotated.

Discussion

Limitations of the algorithm

The system is not valid for those situations where there is not enough functional information (low or null number of homologs with annotated function). An example is shown in Table I (SP:034978). Still, in these cases, a good curve could indicate the presence of a protein family of unknown functionality.

Another limitation is that, due to the elimination of redundant sequences in the input pre-process, functional specificity relying on very subtle sequence differences may be missed (e.g., the inactivation of an active center by a single-point mutation).

Improvements

Further work is needed for the validation of function by complementary analysis of homologs (which could be done by analysis of the sequence similarities of the set of homologs with other members of the database outside the set), and for the inclusion of cases with low number of examples, providing adequate indications of the statistical significance.

Another possibility for development is the automation of the procedure for defining which annotations are synonymous, complementary, hierarchically contained, or contradictory (here done by hand).

An alternative scoring function to the one presented here, could be computed through generation of 'reasonable' amino acid distributions (according to a substitution matrix) and of 'reasonable' versus 'random' matches to it. The analysis of f and s values obtained in such experiments may give insights into a better ϵ scoring function. Another possibility would be to use curated alignments of functional equivalent proteins.

Applications

This system can be the starting point for the improvement of automatic function assignment systems such as GeneQuiz (Scharf *et al.* 1994; Casari *et al.* 1996; Andrade *et al.* 1999). The use of multiple homologs for functional transfer and the transfer of function to fragments of the sequence, are necessary in order to reduce the amount of errors in the annotation process. Another interesting application could be the automatic detection of incoherences in the information already present in public sequence databases.

Acknowledgments

Thanks to Nigel P. Brown for fruitful discussions and for creating and supporting MView and to Joerg Schultz for providing some of the examples.

References

Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; and Lipman, D. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Andrade, M. A., and Sander, C. 1997. Bioinformatics: from genome data to biological knowledge. *Current Opinion in Biotechnology* 8:675–683.

Andrade, M.; Brown, N.; Leroy, C.; Hoersch, S.; de Daruvar, A.; .Reich, C.; Franchini, A.; Tamames, J.; Valencia, A.; Ouzounis, C.; and Sander, C. 1999. Automated genome sequence analysis. *Bioinformatics* 15. In press.

Bairoch, A., and Apweiler, R. 1999. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL in 1999. *Nucleic Acids Res.* 27:49–54.

Bork, P., and Bairoch, A. 1996. Go hunting in sequence databases but watch out for the traps. *Trends in Genetics* 12:425–427.

Bork, P., and Koonin, E. 1998. Predicting functions from protein sequences – where are the bottlenecks. *Nature Genetics* 18:313–318.

Brown, N. P.; Leroy, C.; and Sander, C. 1998. MView: A Web compatible database search or multiple alignment viewer. *Bioinformatics* 14:380–381.

Casari, G.; Ouzounis, C.; Valencia, A.; and Sander, C. 1996. GeneQuiz II: Automatic function assignment for genome sequence analysis. In *1st Annual Pacific Symposium on Biocomputing*, 707–709. Hawaii, USA: World Scientific.

Galperin, M. Y., and Koonin, E. V. 1998. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.* 1:0007. http://www.bioinfo.de/isb/1998/01/0007/.

Koonin, E. V.; Mushegian, A. R.; Galperin, M. Y.; and Walker, D. R. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25:619–637.

Scharf, M.; Schneider, R.; Casari, G.; Bork, P.; Valencia, A.; Ouzounis, C.; and Sander, C. 1994. Genequiz: a workbench for sequence analysis. *Intelligent Systems for Molecular Biology* 2:348–353.

Journal of Molecular Biology
Vol. 297, No. 1, March 2000
ISSN: 0022-2836

Assessing Annotation Transfer for Genomics: Quantifying the Relations between Protein Sequence, Structure and Function through Traditional and Probabilistic Scores

pp. 233-249 (doi:10.1006/jmbi.2000.3550)

Cyrus A. Wilson[1], Julia Kreychman[1], Mark Gerstein[2]

[1]Department of Molecular Biophysics and Biochemistry
[2]Department of Computer Science, Yale University, 266 Whitney Avenue, PO Box 208114, New Haven, CT, 06520, USA

(Received 2 September 1999; received in revised form 5 January 2000; accepted 6 January 2000)

Abstract

Measuring in a quantitative, statistical sense the degree to which structural and functional information can be "transferred" between pairs of related protein sequences at various levels of similarity is an essential prerequisite for robust genome annotation. To this end, we performed pairwise sequence, structure and function comparisons on ~30,000 pairs of protein domains with known structure and function. Our domain pairs, which are constructed according to the SCOP fold classification, range in similarity from just sharing a fold, to being nearly identical. Our results show that traditional scores for sequence and structure similarity have the same basic exponential relationship as observed previously, with structural divergence, measured in RMS, being exponentially related to sequence divergence, measured in percent identity. However, as the scale of our survey is much larger than any previous investigations, our results have greater statistical weight and precision. We have been able to express the relationship of sequence and structure similarity using more "modern scores," such as Smith-Waterman alignment scores and probabilistic *P*-values for both sequence and structure comparison. These modern scores address some of the problems with traditional scores, such as determining a conserved core and correcting for length dependency; they enable us to phrase the sequence-structure relationship in more precise and accurate terms. We found that the basic exponential sequence-structure relationship is very general: the same essential relationship is found in the different secondary-structure classes and is evident in all the scoring schemes. To relate function to sequence and structure we assigned various levels of functional similarity to the domain pairs, based on a simple functional classification scheme. This scheme was constructed by combining and augmenting annotations in the enzyme and fly functional classifications and comparing subsets of these to the *Escherichia coli* and yeast classifications. We found sigmoidal relationships between similarity in function and sequence, with clear thresholds for different levels of functional conservation. For pairs of domains that share the same fold, precise function appears to be conserved down to ~40 % sequence identity, whereas broad functional class is conserved to ~25 %. Interestingly, percent identity is more effective at quantifying functional conservation than the more modern scores (e.g. *P*-values). Results of all the pairwise

comparisons and our combined functional classification scheme for protein structures can be accessed from a web database at http://bioinfo.mbb.yale.edu/align

Keywords: bioinformatics; sequence similarity; percent identity; structure similarity; functional classification

Article

Jump to: Introduction | Discussion and Conclusion
Figure 1 | Figure 2 | Figure 3 | Figure 4 | Figure 5 | Figure 6 | Figure 7 | Table 1

Introduction

The problem of genome annotation

Perhaps the most valuable information to be gained from a genome analysis is functional annotation of all the gene products. Unfortunately, of all the proteins whose sequences are known, functions have been experimentally determined for only a very small number (Andrade & Sander, 1997). Given the current size and accessibility of sequence and structure data, homologs of a newly sequenced gene's product can be identified *via* database searches, and probable structure and function assigned to the gene product (Bork *et al.*, 1998). This is based on the concept that sequence similarity implies structural and functional similarity. However, structural and functional annotations should be transferred with caution. If a protein is assigned an incorrect function in a database, the error could carry over to other proteins for which structure or function is inferred by homology to the errant protein (Brenner, 1999; Karp, 1996, 1998a). In large databases such an error can propagate out of control, presenting a serious quality control issue as we move to larger genomes from multicellular organisms.

Benchmarking fold and function recognition

Here, we used manually curated structural and functional classifications as standards in analyzing to what degree annotations of a protein's structure and function can be transferred to a similar sequence. The knowledge gained from the study can be used to establish confidence levels for structure and function prediction, improving our understanding of how long it will take to annotate accurately an entire genome.

Our simultaneous analysis of relationships between sequence and structure, sequence and function, and structure and function (Figure 1) may provide insight into paradigms for functional prediction other than that based alone on sequence similarity (Enright *et al.*, 1999).

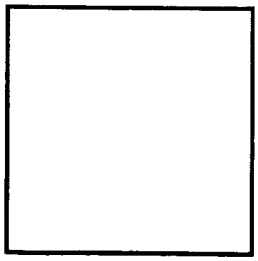


Figure 1 This Figure schematically depicts certain aspects of our comparison methodology. (a) The paradigm relating sequence to structure to function. There has not been as much assessment of functional annotation transfer based on structure as there has been with sequence-based structural and functional annotation transfer. (b) How we conceptualized our analysis in terms of pairs. A few examples of SCOP domains (identified on the left and bottom) are included from our comparison. In the Figure the shape represents fold, and the pattern represents function. We have highlighted some example categories of pairs: a pair that shares fold and function, a pair that shares fold but not function and a pair that shares neither fold nor function. The latter category of pairs is not considered in our investigation; we looked only at paired domains with the same fold. In constructing our pairs, we used only a representative set of SCOP domains. This is illustrated in the Figure by the domains flagged with asterisks. Note, in

particular, that the SCOP domain d4tima_ is not paired with anything because it is represented by d5tima_, which is the same species and protein. For each level of pairs (fold, superfamily, family), cluster representatives were chosen for the level below: (i) for family pairs, one representative was selected from each species/protein, the level below, and then paired with all the other representatives within its family; (ii) for superfamily pairs, one representative was chosen from each family, unless there were domains in the family that shared less than 40 % sequence identity, in which case additional representatives were included, each not more than 40 % identical with the other representatives from the family (this occurs, for instance, for the globins); and (iii) likewise for fold pairs, one representative was chosen from each superfamily, more if there were domains with less than 40 % sequence identity. (c) Subdivides the pairs into the four SCOP classes from which they were composed: (i) all-α, domains consisting of α-helices; (ii) all-β, domains consisting of β-sheets; (iii) α/β, domains with integrated α-helices and β-strands; and (iv) $\alpha+\beta$, domains with segregated α-helices and β-strands. We initially set apart the immunoglobulins from the rest of the all-β pairs because we realized that their large number biases our data. However, we compared the results for the immunoglobulin pairs to all other pairs and found that they generally exhibit the same behavior as the other pairs. Therefore we decided to leave them in the comparison.

Past results

Sequence-structure

The transfer of structural annotation is well characterized. Chothia & Lesk (1986, 1987) found that structural divergence, when expressed in terms of the RMS separation of matching alpha carbon atoms, was an exponential function of sequence divergence, expressed in terms of the fraction of residues that differed between sequences. The reliability of structural annotation transferred by homology, then, depends on the sequence identity of the homologous proteins (Chothia & Lesk, 1986). Flores *et al.* (1993), Russell & Barton (1994), and Russell *et al.* (1997) observed the same general trend, and also characterized the conservation of structural features other than the C$^{\alpha}$ backbone, such as secondary structure, accessibility and torsion angles. A paper by Wood & Pearson (1999) re-expressed the sequence-structure relationship in terms of statistically based "Z-scores" and found that this relationship had a simple linear form in terms of these scores. They also noted that protein families differed in detail in the slope of this linear relationship.

Others have focused on the limits of sequence comparison, specifically around the "twilight zone," the region of sequence similarity that does not reliably imply structural homology (Doolittle, 1987), and on establishing cut-offs for significant sequence similarity. Using the SCOP structural classification (Murzin *et al.*, 1995), Brenner *et al.* (1998) benchmarked the effectiveness of the popular FASTA and BLASTP programs and their probabilistic scoring schemes (i.e. the *e*-value) (Pearson & Lipman, 1988; Pearson, 1996; Altschul *et al.*, 1990, 1994; Karlin & Altschul, 1993). They found that in making fold assignments, the FASTA *e*-value closely tracked the number of false positives, i.e. the error rate, and that at a conservative *e*-value cut-off of 0.001, the FASTA program could detect nearly all the relationships that would be detected by a full Smith-Waterman comparison (Smith & Waterman, 1981). Specifically, they found that FASTA with a 0.001 threshold would find 16 % more of the structural relationships in SCOP than would be found by standard sequence comparison with a 40 % identity threshold. This rigorous benchmarking approach has been extended to assess transitive sequence comparison, through a third intermediate sequence and multiple-sequence matching programs such as PSI-blast (Park *et al.*, 1997, 1998; Gerstein, 1998a; Salamov *et al.*, 1999). In a related study Rost (1999) worked on characterizing

the region after the twilight zone, which he called the "midnight zone". In a sense these benchmarking studies have culminated in the CASP fold recognition experiments (Moult *et al.*, 1997; Stemberg *et al.*, 1999).

Sequence-function

Although the exact dependence of functional similarity on sequence and structural similarity is not completely clear, initial indications of a gene product's function are most often based on simple sequence similarity (Bork *et al.* 1994, 1998). Often these are merely based on the best hit in database comparisons; see, for example, the annotation of some of the early genomes (Fraser *et al.*, 1995, 1998). However, possibilities for more robust annotation transfer are increasingly available. One looks at the pattern of hits amongst different phylogenetic groups (Tatusov *et al.*, 1997). Often these focus on the existence of key motifs and patterns associated with function (Zhang *et al.*, 1998; Bork & Koonin, 1996; Attwood *et al.*, 1999).

Sequence-structure-function

One way that the better-defined sequence-structure relationship can assist in function prediction is initially to predict the structure of an uncharacterized sequence and then predict the function based on the limited repertoire of functions known to occur with that structure. To some degree this was achieved by Fetrow and co-workers (Fetrow *et al.*, 1998; Fetrow & Skolnick, 1998). They predicted structural profiles based on threading and *ab initio* methods, and then searched with these against profiles of known structures in order to predict function.

In related work, Russell *et al.* (1998) discussed using identification of structural binding sites in predicting protein function. In a comprehensive study, Hegyi & Gerstein (1999) investigated to what degree folds were associated with functions. They found that most folds were associated with one or two functions with the exception of a few special folds, such as the TIM barrel, that could carry out numerous functions. Furthermore, they found that particular folds were often confined to distinct phylogenetic groups, an additional fact that can feed into an integrated sequence-structure-function analysis (Gerstein & Hegyi, 1998; Gerstein, 1997, 1998b,c).

Here, we look at pairwise comparisons of protein sequence, structure and function among proteins that share the same fold. We assess the trends relating sequence, structure and function and consider the implications for structural and functional annotation transfer.

New developments: probabilistic scoring and growth of the databank

The past studies regarding sequence, structure and function relationships often used RMS separation and percent sequence identity (or a linear variant of it, such as the fraction of mutated residues) to express similarities in structure and in sequence, respectively. However, it has become increasingly common to use probabilistic scoring schemes (*P*-values) to express the quality of a match in terms of statistical significance rather than an arbitrary raw score such as percent identity (Pearson, 1998; Karlin & Altschul, 1990, 1993; Karlin *et al.* 1991; Altschul *et al.* 1994; Bryant & Altschul, 1995; Abagyan & Batalyov, 1997). With *P*-values, scores from different investigations can be compared in a common framework. Recently, it was found that sequence and structure similarity significance can be expressed as *P*-values in the same unified statistical framework (Levitt & Gerstein, 1998). Here, we use such probabilistic scoring methods to overcome the limitations of the more traditional scores.

Another recent development is the tremendous growth in the number of solved structures. The RCSB Protein Data Bank (Bernstein *et al.* 1977) now contains more than 10,000 protein structures. These structures are broken into more than 18,000 domains, and then domains that share a fold are paired up with each other for comparison (Figure 1(b)). Here, we

survey ~30,000 pairs of protein domains that are known to have the same fold, approximately 1000 times the number compared by Chothia & Lesk (1986). The large scale of this comparison affords greater statistical weight to the results.

Alignment of 30,000 pairs from SCOP

The basic unit of comparison: a pair of protein domains

The protein domains that we studied were classified by SCOP, a Structural Classification of Proteins (Murzin *et al.* 1995; Brenner *et al.* 1996; Hubbard *et al.* 1997), a hierarchy of five levels: (i) class, domains that have the same secondary structural content (all-α, all-β, α/β, or $\alpha+\beta$); (ii) fold, domains that geometrically share the same tertiary fold; (iii) superfamily, domains descended from the same ancestor (but which lack measurable sequence similarity); (iv) family, domains in the same protein sequence family (which have appreciable sequence similarity); and (v) species and protein.

Pairs of protein domains that are grouped together at the fold, superfamily or family level form the basic unit of our comparisons.

Selection of pairs

There is potentially a huge number of pairs of domains that can be constructed out of the relationships in SCOP. For instance, in the current version of SCOP there are ~3.9 million potential pairs between domains sharing the same fold. Most of these are between nearly identical structures. In order to keep the number of pairs manageable, we used a straightforward clustering scheme, described in the legend to Figure 1. We selected 29,454 representative pairs from the total in SCOP. To achieve a wide range of similarities, we constructed the pairs on three levels of the SCOP hierarchy: (i) family pairs, 19,542 pairs of domains in the same family; (ii) superfamily pairs, 4220 pairs of domains in the same superfamily but different families; and (iii) fold pairs, 5692 pairs of domains in the same fold but different superfamilies.

All the selected domains were at least 50 residues in length and were drawn from the four major SCOP secondary-structural classes: all-α, all-β, α/β, and $\alpha+\beta$ (Figure 1(c)).

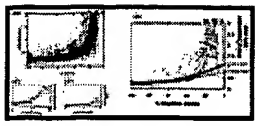
We automatically aligned each of our selected domain pairs twice, once by global Needleman-Wunsch sequence comparison (Needleman & Wunsch, 1971; Myers & Miller, 1998) and then by structure (Gerstein & Levitt, 1996, 1998), calculating scores for sequence and structural similarity.

Web-accessible database

The results of all the pairwise comparisons are available *via* a searchable database on the web at http://bioinfo.mbb.yale.edu/align The query engine allows searches of individual SCOP pairs, all pairs that include a given SCOP domain, or all pairs containing any SCOP domain contained in a given PDB entry.

Traditional scores: RMS and percent identity

The sequence-structure relation, as expressed by the root-mean-square (RMS) of the aligned C^{α} distances and percent sequence identity, has been previously characterized as an exponential function by Chothia & Lesk (1986) and others (Flores *et al.* 1993; Russell & Barton, 1994; Russell *et al.* 1997). As Figure 2 illustrates, our data display a similar trend. (Exact equations are given in the legend to Figure 2.) However, we have one thousand times as many data points as in Chothia and Lesk's original study (30,000 as opposed to 30).



(larger image: 22K)

Figure 2 RMS as a function of percent identity. (a) A simple scatter plot of our pairs, relating RMS separation to percent sequence identity. This is similar to the presentation given by Chothia & Lesk (1986), but in this survey we looked at 30,000 pairs, 1000 times the number they compared. Outliers (pairs with RMS scores further than two standard deviations from the mean for their percent identity) are excluded from this graph; they represent domains that are very closely related with the exception of a conformational change. (b) A simplified graph with a number of fits to the data. For each percent identity bin we show the median RMS value, indicated by (◊) and the top and bottom quartile RMS values, indicated by the bars. Two fits are drawn through the median RMS values. The thin line, labeled SINGLE, is a simple exponential fit through the medians. It has the form:

$$R = 0.21e^{0.0132H}$$

where R is the RMS deviation after least-square fitting, H is the percent difference between the sequences (H for Hamming distance), and H=100 %-I, where I is the percent sequence identity. The thick line, labeled MULTI, is a multigraph fit, which is described in the legend to Figure 4. The relation between RMS and percent identity according to this fit is expressed by the equation:

$$R = 0.18e^{0.0187H}$$

The twilight zone of sequence identity and below is labeled TZ. In this region, sequence similarity is not significant and not reliable for predicting structural similarity. This is why the median values in this area of the graph deviate significantly from the fits, which consider only data above 20 % sequence identity. For reference we include the original data points from Chothia and Lesk's, 1986 paper (A.M. Lesk, personal communication), indicated by X. Their data follow the form:

$$R = 0.40e^{0.0187H}$$

The difference between the Chothia & Lesk trend and our relationship is due to the different trimming methods used in calculating the RMS score. Chothia and Lesk imposed a 3 Å cut-off in determining the conserved core residues; we defined the core as the better matching (in terms of C^{α} distances) half (50 %) of the residue pairs. (c) and (d) The effect our trimming has on median RMS values. The RMS values in (c) are calculated from all the matched residues in each pair; the values in (d) are calculated from the better matching 50 % of the residues.

The main difference between our results and the previous studies is due to differences in RMS "trimming" methods. By trimming we refer to the process of removing the worst-fitting aligned atoms from the RMS calculation, to arrive at a structural "core." This was first developed in Lesk's sieve-fit procedure (Lesk & Chothia, 1984) and has been refined in numerous studies (e.g. Gerstein & Altman (1995)). This is done because the small distances between well-matched alpha

carbon atoms have much less of an effect on the RMS than do the very large distances between poorly matched atoms. The untrimmed score of divergent protein domains is then concerned primarily with the poorly matched residues instead of the conserved core. Trimming alleviates this effect by restricting the RMS calculation to include only those residues believed to be in the conserved core. However, the degree of trimming is to some extent arbitrary, and this choice affects the baseline of the reported RMS scores. Here we considered only the better half (50 %) of matched residues in a given pair of protein domains. Chothia & Lesk (1986) chose a somewhat different threshold. Figure 2(c) and (d) demonstrate the effect of trimming.

(larger image: 1K)

Figure 3 Similarity scores: structural comparison score as a function of Smith-Waterman score. Alignment similarity scores S_{str} and S_{seq} have certain advantages over RMS and percent identity scores for expressing the sequence-structure relation. S_{str} is calculated according to equation (1) in the text (Gerstein & Levitt, 1998; Levitt & Gerstein, 1998). S_{seq} is calculated using the BLOSUM50 matrix (Henikoff & Henikoff, 1992) with gap opening and extension penalties of -12 and -2, respectively. (a) This is analogous to (b) in Figure 2. From the original 30,000 pairs we show the median S_{str} value for each S_{seq} bin, along with quartile bars above and below. Again the twilight zone and below is labeled TZ. The thin line, marked SINGLE, is a simple fit to the median S_{str} values in this graph; it has the form:

$$S_{str} = 2144 - 1106\exp(-0.00544 S_{seq})$$

The thick fit, marked MULTI, is the multigraph fit, explained below. It follows the equation:

$$S_{str} = 2157 - 787\exp(-0.0028 S_{seq})$$

The equations presented here provide an approximation of the observed trends; as (b) illustrates, they are nothing more than simple approximations. The main disadvantage of S_{str} as a measure of structural similarity is its heavy length dependency for pairs of structurally similar protein domains. (b) Surface plot of the median S_{str} as a function of S_{seq} and alignment length (the number of matched residue pairs). It is clear that the size of the aligned domains plays a major role in the resulting S_{str}, even though our fits do not take length into account. (c) and (d) Relate S_{seq} and S_{str} to the more familiar percent identity and RMS measures. The fits were used to convert between scoring schemes in constructing the multigraph fit. We derived the multigraph fit in order to create one set of equations and parameters that would relate sequence and structural similarity using either the percent identity and RMS scheme or the S_{seq} and S_{str} scheme, and allow translation between them. We simultaneously performed least-squares fits to the median values in four graphs: Figures 2(b) and 3(a) and the calibrations of S_{seq} to percent identity and S_{str} to RMS, (c) and (d), respectively. In all cases, we ignored data in and below the sequence identity twilight zone (labeled TZ). The parameters in (a) are dependent on the parameters in Figure 2(b)*via* the mentioned calibrations.

Analogous alignment similarity scores: Smith-Waterman score and structural comparison score

The dependence of the RMS separation on trimming method restricts its usefulness in comparing data. Likewise, there are many problems with using percent identity as a measure of sequence similarity. For instance, a match of non-identical but still similar residues (e.g. Arg *versus* Lys) scores the same as one between completely different residues (e.g. Arg *versus* Val), and gaps do not enter in the score calculation. Consequently, we now turn to alignment similarity scores, which eliminate some of the problems with traditional scores.

For sequence alignments, an alignment score is defined as the sum of the similarity matrix values for the alignment, minus the total gap penalty. This is sometimes called the Smith-Waterman score (Smith & Waterman, 1981). An analogous alignment score for structure is the structural comparison score, described by Levitt & Gerstein (1998). We will refer to these two similarity scores as S_{seq} and S_{str}, respectively. Note that they both increase for more similar pairs, whereas RMS increases for more divergent pairs. Specifically, S_{str} is the score maximized by the structural alignment program we used (Gerstein & Levitt, 1998). It can be calculated from any pair of aligned structures according to the function:

$$S_{str} = M \sum \left(\frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} - \frac{N_{gap}}{2} \right) \quad (1)$$

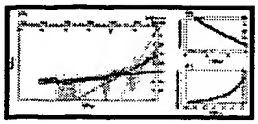
M and d_0 are constants, usually set to 10 and 5 Å, N_{gap} is the number of gaps in the alignment, d_i is the distance between each aligned pair of Cα atoms, and the sum is carried over all aligned pairs, i.

The main advantage of S_{str} over RMS in describing structural similarity is that the Cα to Cα distance, d_i, appears in the denominator of the calculation. This means that the smallest distances, corresponding to the best matches in the conserved core, are most significant in determining the score. Hence, the need for trimming is eliminated. S_{str} is also advantageous because it takes gaps into account and because of the fundamental analogy between this score and S_{seq}.

Figure 3(a) displays the relationship between structural and sequence similarity as expressed by S_{str} and S_{seq}. Figure 3(c) and (d) show calibration curves relating each of these scores back to approximate RMS separation and percent identity, respectively. Calibration curves help one get an intuitive feel for the degree of relationship in terms of the more traditional scores. Figure 3(b) adds a third axis, alignment length, and demonstrates that S_{str} depends greatly on this quantity. Although S_{str} and S_{seq} are "better" scores than RMS and percent sequence identity, the heavy dependence of both of these on length limits their usefulness in many situations. In other words, two pairs of similar domains with equal percent sequence identities but different lengths can have drastically different S_{seq} scores.

Probabilistic scores: P-values expressing the significance of sequence and structure similarity

Probabilistic scores can, to a great degree, overcome the length-dependence problems associated with the alignment scores. Probabilistic measures are advantageous because they express similarity not by an arbitrary "score" but by a statistical significance: the likelihood that such a similarity could be achieved by chance. This likelihood is also called the "*P*-value." We used calculations (described in detail in the legend to Figure 4) based on those given by Levitt & Gerstein (1998) to obtain *P*-values based directly on S_{str} and S_{seq}; we refer to these calculated *P*-values as P_{str} and P_{seq}, respectively. For P_{seq} we could equally well have used the numbers from one of the popular sequence search programs (i.e. BLAST or FASTA) as all these values have been shown to be perfectly proportional to each other (Levitt & Gerstein, 1998; Brenner *et al.* 1998).



(larger image: 19K)

Figure 4 Probabilistic scores: *P*-values. P_{seq} and P_{str} are *P*-values calculated from S_{seq} and S_{str} according to the formalism given by Levitt & Gerstein (1998). Both quantities have the same overall functional form in terms of an extreme value distribution:

$$P = 1 - \exp(-\exp(-Z))$$

where *P* is either P_{seq} or P_{str}. For P_{seq}, $Z=S_{seq}/a-2 \ln M-b/a$, where a=5.84, b=-26.3, and *M* is the geometric mean of the lengths of the two sequences (i.e. $M^2=nm$, where *n* and *m* are the two sequence lengths). For P_{str}, *Z* is a function of S_{str} and *N*, the number of matched residues: For $N<120$:

$$Z = (S_{str} - c\ln^2 N - d\ln N - e)/(f\ln N + g)$$

For $N \geq 120$:

$$Z = (S_{str} - a\ln N - b)/(f\ln 120 + g)$$

At N=120, continuity implies that:

$$a\ln 120 + b = c\ln^2 120 + d\ln 120 + e \quad \text{and} \quad a = 2c\ln 120 + d$$

This, in turn, allows the calculation of the constants:

$$a = 171.8,\ b = -419.4,\ c = 18.4,\ d = -4.50,\ e = 2.64,\ f = 21.4,\ g = -37.5$$

(a) of this Figure is analogous to Figures 3(a) and 2(b), with the exception of the fits. It is a log-log (base 10) plot relating P_{seq} and P_{str}. We show the median $\log(P_{str})$ value for each $\log(P_{seq})$ bin, along with quartile bars above and below. We have added approximate percent identity and RMS values to the *x* and *y* axes to aid interpretation of the graph in terms of more familiar scores. The values were calculated using the calibration curves in (b) and (c). The straight-line nature of the log-log plot reveals distinct relations inside and outside the twilight zone, labeled TZ. (The area of percent identity below the twilight zone does not appear in P_{seq} graphs, there is no significance for such low sequence similarity; thus all data points in that zone appear at P_{seq}=1 or $\log[P_{seq}]$=0.) The thick line in the figure is fit to the median P_{str} values for P_{seq} values outside the twilight zone; its equation is:

$$P_{str} = 10^{-10} P_{seq}^{0.05}$$

The thin line is fit to the data inside the twilight zone; it follows the relation:

$$P_{str} = 10^{-6} P_{seq}^{0.274}$$

For reference we include the dotted line, representing the function $P_{str}=P_{seq}$, where sequence and structural similarity are equally significant. See the text for a discussion of how the two trends might be interpreted with respect to this line.

P_{seq} and P_{str} can be used to express the relationship between structure and sequence similarity on a more fundamental level. Figure 4(a) shows a log-log (base 10) plot of P_{str} against P_{seq}. Because it is log-log, trends can be visualized as straight lines. Two straight lines are necessary to fit the points well, with the discontinuous boundary between the lines located at the beginning of the twilight zone. The different slope of the line at low sequence similarity reveals that in the twilight zone there is a different relationship between the significance of structural similarity and that of sequence similarity. In particular, for domain pairs in the twilight zone (according to the percent identity to P_{seq} calibration in Figure 4(b)), structural similarity is more significant than sequence similarity (having a smaller *P*-value or more negative log *P*-value). In contrast, for pairs with more than ~30 % identity, the situation is reversed, with a given pair having more significant sequence similarity than structural similarity. One possible interpretation of this reversal is as follows. Structure is always more highly conserved than sequence, so usually a given amount of structural similarity is not as significant as a corresponding amount of sequence similarity. However, this is true only when meaningful sequence similarity actually exists; thus, it does not apply in the twilight zone, where sequence similarity is by definition not significant. Note that all pairs in our comparison share at least the same fold, implying that they always have a significant amount of structural similarity.

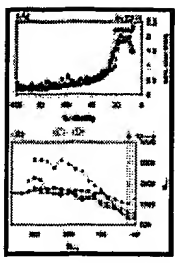
In other words, for closely related sequences, differences in sequence similarity are more meaningful, whereas for highly diverged sequences that share the same fold, the differences in structural similarity are more significant.

Fitting two lines to the P_{str}*versus*P_{seq} graph suggests that the same might be done for other scoring schemes. It is possible to some degree to fit the traditional RMS *versus* percent identity graph (Figure 2) with two straight lines instead of an exponential cruve. However, in this case, we opted for the more conventional presentation.

Class differences

The division of SCOP into classes based on secondary-structural composition allows easy investigation as to whether there are any deviations from the common similarity relationships on account of secondary-structure characteristics. Figure 5(a) reveals that secondary structural composition does not markedly affect the trends in sequence and structure similarities. This is consistent with the data given by Wood & Pearson (1999). However, the larger average length of α/β domains compared with domains in the other classes results in a deviation in the length-dependent S_{str} (Figure 5(b)). The consistency among length-independent scores applies for certain individual folds as well. The immunoglobulin fold makes up an appreciable fraction of all the β-pairs (Figure 1(c)), yet the results are not affected if these pairs are left out.

Figure 5 SCOP class differences. Previously it has been observed that secondary structural composition does not cause deviations from the trends in structure and sequence similarity (Flores *et al.* 1993). To test this observation we looked at the scores divided by SCOP class. The following legend applies to the graphs: (-■-), all alpha; (- ◊ -), all beta; (- - ▲ - -), alpha/beta; (- -x- -), alpha+beta. (a) Median RMS values for each percent identity bin. The traditional scores reveal no dependency on class. However, in (b) α/β pairs consistently score



(larger image: 13K)

higher S_{str} scores than pairs in other classes. This is a consequence of the dependence of S_{str} on length; domains in the α/β class are longer, on average, than in the other classes.

Linking sequence and structure to function

Difficulties of functional comparison

There is a clear, well-characterized relationship between sequence and structure similarity, which can be used to transfer precisely structural annotation based on the degree of sequence homology. In genome analysis, however, one is usually more interested in finding a functional annotation for an open reading frame based on similarity to well-known proteins; yet the sequence-function and structure-function relationships have not been as explicitly characterized. The fundamental obstacle to extending this and similar investigations to deal with function is the absence of a clear measure of functional similarity. Although we were able to present three different quantitative measures of structural relatedness, an analogous situation for function does not exist. How can one express quantitatively the degree of similarity between a triosephosphate isomerase and a glucose-6-phosphate isomerase? How do they compare to trp repressor?

The absence of a clear measure of functional similarity is not the only obstacle in transferring the functional annotations between proteins with different degrees of homology. The definition of function itself is often vague. More specifically, at present there is an absence of such important information as a standardized vocabulary for protein functional annotations with an associated numbering scheme, descriptions of monomer functions of subunits of multisubunit proteins and hierarchical functional assignments for proteins with multiple functions. As a consequence of these difficulties there is no functional equivalent to the hierarchical fold classification for domains in PDB.

As signs of progress in this direction, several functional classifications have been developed to date. One is the ENZYME system developed by the Enzyme Commission (EC) to classify enzymes by reaction type (Webb, 1992). This system has the advantage that it is "universal," applicable to proteins in many different organisms, and is in wide use. However, it also has several drawbacks. First of all, it does not consider catalytic reaction mechanisms (Riley, 1998a), often ignoring obvious similarities. Second, it presumes a 1:1:1 relationship between gene, protein and reaction, although this is often not the case (an enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function). Perhaps the most significant drawback of the EC classification is that it applies to only enzymes.

A number of more comprehensive schemes have been developed, which classify non-enzymes as well as enzymes. Most of these focus on individual organisms. Several such schemes exist, for instance, GenProtEC/EcoCyc for *E. coli* (Karp *et al.*, 1998b; Riley & Labedan, 1996; Riley, 1998b), MIPS for yeast (Mewes *et al.*, 1998), Ashburner's functional classification for Drosophila, which is connected to FLYBASE (Ashburner & Drysdale, 1994), and EGAD for human ESTs (Adams *et al*, 1995). These classifications possess some advantages. They have additional levels of hierarchy that help present a more comprehensive picture of genotype-phenotype relationships. On the other hand, these classifications still leave much room for improvement. For example, there is no standardized vocabulary to allow for keyword searches among multiple databases and across organisms, and there are inconsistencies in category numbering style.

Finally, there has been some promising work going beyond the ENZYME and organism-focused classifications. There has been progress on completely automated functional classification (des Jardins *et al.*, 1997; Tamames *et al.*, 1997), which has the potential for putting function assignments on a more objective basis. There are a number of databases synthesizing the various enzyme functions into coherent pathways and systems (e.g. KEGG and WIT, Ogata *et al.*, 1999; Selkov *et al.*, 1998). There also have been some very recent attempts to develop cross-species classifications of non-enzyme functions in the framework of the Gene Ontology Project (GO, geneontology.org). GO is a joint project between FlyBase, the Saccharomyces Genome Database and Mouse Genome Informatics, attempting to merge the fly, yeast and mouse functional classification schemes. However, a truly universal system for classifying all protein functions in all organisms within the same framework remains quite a challenge because of the sheer diversity of organisms and distinct protein functions.

Our simple functional classification of SCOP domains: FLY+ENZYME

Given the discussed limitations, we constructed a simple functional classification for the SCOP domains included in our comparison; our classification is based on a merger of two of the existing functional annotations and a cross-referencing of subsets of this combination with some of the organism-specific schemes. First, we used pairwise comparison to cross-reference the PDB domains against the Swissprot database (Bairoch & Apweiler, 1998), as described by Hegyi & Gerstein (1999). We chose to assign protein functions according to Swissprot because it provides more comprehensive functional annotations than SCOP.

We were initially able to divide all entries into enzymes and non-enzymes, a division that represents the highest level of functional difference in our classification scheme (Figure 6). For the enzyme category, we transferred EC (Webb, 1992) numbers to those SCOP domains with a one-to-one match to a Swissprot enzyme. Only one-to-one matching entries could be considered because Swissprot assigns ENZYME numbers to entire proteins, whereas SCOP is a domain-based classification; therefore we could be confident about the classification of only those domains which map to an entire Swissprot entry.

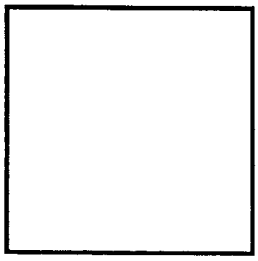


Figure 6 Functional classification of enzymes and non-enzymes. (a) Divides the pairs by general function. There are three categories of pairs: (i) enzymes paired with non-enzymes (no general functional similarity), labeled ENZ/~ENZ; (ii) enzymes paired with enzymes (same general function), labeled ENZ/ENZ; and (iii) non-enzymes paired with non-enzymes (same general function). Pairs for which one or both domains could not be identified as enzyme or non-enzyme are not included in this chart. Enzymes are classified according to the EC system (Webb, 1992). The first component of the number represents the nature of reaction and is called class. There are six classes: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The next level is subclass. It refers to the chemical groups on which the enzyme acts. For example, the first class, oxidoreductases, has 19 subclasses that are arranged according to the donor group that undergoes oxidation (CH-OH, aldehyde or oxo group, CH-CH group, etc). For another group of enzymes (hydrolases) subclass is determined by the nature of the bond: ester bond, peptide bond, etc. The next level is sub-subclass. For oxidoreductases this indicates the acceptor group: NAD(+) and NADP(+), or cytochrome; for hydrolases the sub-subclass represents the nature of substrate (carboxylic ester hydrolases, thiolester hydrolases, etc.). The fourth level represents a unique number for each individual enzyme, for example, 1.1.1.1: alcohol dehydrogenase. (b) Shows how we

adapted the functional classification of Drosophila gene products developed by M. Ashburner. This classification is loosely connected with FLYBASE (Ashburner & Drysdale, 1994). We used version 1.55 (4 August 1997) that was available from Ashburner's website:

http : //www.ebi.ac.uk/ ~ ashburn

The specific files that we used were taken from the ftp directory:

ftp.ebi.ac.uk/databases/edgp/misc/ashburner

We refer to these as constituting the original FLY classification. Recently, the FLY classification has been superceded by the GO (Gene Ontology) Project classification, which merges fly, mouse and yeast annotation. Files related to the GO classification are available from www.geneontology.org In the original FLY classification all members of the highest level are labeled 0, representatives of the next level are labeled 1, and all lower levels are labeled 2 through to 9. We changed the numbering scheme so that it will reflect the hierarchical nature of the classification. This Figure illustrates sections of the original and modified classification. The top level in the FLY classification scheme is called "Function primitive" (level 0) and includes five classes: "Metabolism,""Intracellular protein traffic,""Cell structure,""Developmental process,""Physiological process," and "Behavior." The next level after "Function primitive" is "Process" or "Molecule" (level 1 in Ashburner's classification). For "Function primitive - Metabolism" the processes are "Carbohydrate metabolism,""Nucleotides and nucleic acids metabolism," etc. For "Function primitive - Cell Structure" the "Process" can be "Nucleus,""Mitochondrion,""Membrane," etc. The next level is "Pathway" or "Macromolecule" (level 2 in the original classification). "Pathway" can include "Metabolic pathway,""Signaling pathway," or "Developmental pathway." The "Macromolecule" category includes "Protein" and "Nucleic Acid". We added categories to the original classification in order to classify some mammalian proteins that are widely represented in SCOP but are absent from the original FLY scheme. These categories include immune system proteins (labeled "new" in (b) and respiratory proteins such as hemoglobin and myoglobin that we added to "Function primitive - Physiological process - Respiration". We call our adaptation of the original FLY scheme, FLY+. Further information on this adaptation is available at:

http : //bioinfo.mbb.yale.edu/align/func

(c) The overall hierarchy of our final scheme and identification of the different levels of similarity. If two proteins are both enzymes or both non-enzymes, then they possess general functional similarity. If they share the first component of their classification numbers, then they are in the same functional class. If they share the first three components of their enzyme numbers (or the equivalent for non-enzyme numbers, depending on category) then they have the same precise function. A significant

difference between the two main branches of the hierarchy is that the levels of the ENZYME classification do not correspond exactly to those in the FLY+ system because the fly classification is more extensive than the enzyme classification. For instance, the FLY classification takes into account aspects of cellular (cytoskeleton, metabolic pathways, etc.) and phenotypic function (morphology, physiology, behavior) that are absent from the ENZYME scheme. This makes our classification of SCOP proteins somewhat unbalanced, as non-enzymes have much broader and more loosely defined functional classes. As a consequence, while each enzyme is assigned a four-component number, the length of a non-enzyme number varies, depending on the functional category to which it belongs. For example, myosin is assigned a number that happens to have the same length as EC numbers: 3.12.1.1. However, transcription factors are numbered 1.12.9.1.1.1. We took into account this varying hierarchy depth in deciding how many components are necessary to identify precise function in each category. Note that what we mean by domains having the same precise function is not the same as the domains coming from the same essential protein.

In the absence of an EC-type classification for non-enzymes, we assigned functions to non-enzymatic SCOP domains according to Ashburner's original classification of Drosophila protein functions. This classification is derived from a controlled vocabulary of fly terms. It is available on the web and loosely connected with the FLYBASE database (Ashburner & Drysdale, 1994). For clarity, we precisely describe the specific files and version (1.55, 1997) of the classification that we used in the caption to Figure 6, and we will hereafter refer to these data files as constituting the original FLY classification.

The FLY classification is a dynamic object, changing as more is learned about the fly and other organisms. This is particularly true of late with the imminent completion of the Drosophila genome. In fact, since the completion of our analysis, the FLY classification has been superceded by the new GO classification (see above).

The hierarchical structure of the FLY classification makes it well suited for classifying non-enzymatic SCOP entries in a manner comparable to the ENZYME assignments for the enzymes. Another advantage of this classification is that it is more compatible with the makeup of the PDB than the *E. coli* and yeast classifications, as Drosophila is a multi-cellular organism, and many of the known structures come from animals. We were able to use the original FLY classification as a framework to which we added functional categories and individual proteins. For instance, we added "Hemoglobin" to the "Physiological Processes - Respiration" category. Another example is the "Physiological processes - Immunity" category (Figure 6(b)), to which we added immune system proteins. Many of the additions would not be necessary in the context of the new cross-species GO system. We also modified slightly the numbering scheme in the original FLY classification in order to assign a unique hierarchical number to each protein domain (Figure 6(b)). We will refer to our augmented FLY classification as the FLY+ scheme, and our merged scheme as the FLY+ ENZYME classification.

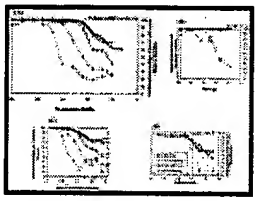
As discussed earlier, the universal functional classification of proteins is very challenging and may not be possible with the current level of knowledge about genes, proteins and genomes. Consequently, the FLY+ENZYME classification of SCOP proteins is somewhat incomplete and inconsistent and retains many of the limitations of its components (Hegyi & Gerstein, 1999; Riley, 1998a). It is not yet broad enough to include many plant, virus and bacterial proteins. Nevertheless, it was sufficient for our analysis, as we were able to classify a very large number of the total 30,000 pairs.

Determining functional similarity

Using our compound functional classification, we were able to assign a level of functional similarity to each domain pair. According to our scheme, a pair can have no functional similarity (an enzyme paired with a non-enzyme) or it can have one of three levels of similarity:

- General similarity. Both domains are enzymes or both are non-enzymes.
- Same functional class. Both domains share the first component of their ENZYME or FLY+ numbers, e.g. 1.1.1.1 alcohol dehydrogenase and 1.3.1.1 cortisone beta-reductase (for enzymes), or 3.3.2.1.2 calcicyclin and 3.6.3.2.1 calmodulin (for non-enzymes).
- Same precise function. Both domains share three components of their ENZYME or FLY+ number, e.g. 1.1.1.1 alcohol dehydrogenase and 1.1.1.3 homoserine dehydrogenase (for enzymes) or 1.2.9.1.1.1 Arc repressor and 1.2.9.1.1.1 C-jun (for non-enzymes; both are transcription factors). A pair that shares precise function must also, by definition, share functional class and general similarity.

Based on those assignments we calculated the percentage of total pairs at a given level of sequence or structural similarity possessing each level of functional similarity. The results appear in Figure 7.



(larger image: 25K)

Figure 7 Linking sequence, structure and function. We express functional similarity as the fractional percentage of pairs at a given level of sequence/structural similarity for which the paired domains share a precise function, functional class, or general similarity (according to our classification, see Figure 6). The following legend applies to (a) through (c): (-o-), general similarity; (-×-), non-enzymes with same functional class; (-▲-), enzymes with same functional calss; (- - -×- - -), non-enzymes with same precise function; and (- - -▲- - -), enzymes with the same precise function. (a) Relates functional similarity to sequence similarity in terms of percent identity. The functional similarity appears as a sharp sigmoid, with distinct thresholds of divergence for precise function, functional class, and general similarity. Enzymes are paired with non-enzymes only at very low percent identity, in and below the twilight zone (labeled TZ). At slightly higher sequence identity, pairs diverge with respect to functional class, and beyond 40 % identity with respect to precise function. Note that 50-100 % identity is not shown because almost all domains that are that similar share function with their counterparts. (b) Shows the same data using P_{seq} as the measure of sequence similarity. Only the divergence in precise function is visible because there is such little significance for the low sequence similarity at which functional class and general similarity diverge, all data points in that region appear near P_{seq}=1 or log[P_{seq}]=0 (the *y*-axis). (c) Illustrates that the structure-function relation is not as clearly defined as that for sequence and function. Functional similarity expressed in terms of RMS separation appears as a broad sigmoid curve; there are thresholds of divergence for precise function, but the divergences in functional class and general similarity are more gradual. The thresholds are apparent only because RMS clusters the most structurally similar pairs between scores of 0 and 0.5 Å. For this reason, RMS is better at discerning functional similarity than S_{str} and P_{str}, which do not cluster the most similar pairs around a set limit. (d) Shows the same relationships (functional conservation *versus* percent identity) as in (a), except that for this graph functional similarity is determined in terms of the MIPS (Mewes *et al.*, 1998) and GenProtEC (Riley, 1998b) classifications rather than the FLY+ENZYME scheme. The legend appears as the inset

on the graph. We assigned MIPS and GenProtEC classifications to SCOP domains based on sequence comparisons to classified yeast and *E. coli* open reading frames (ORFs), respectively. The SCOP domain most closely matching each ORF classified in MIPS or GenProtEC was assigned the corresponding MIPS or GenProtEC function number. Only matches of 80 % sequence identity or greater were considered. We used this SCOP domain as a functional representative; when determining functional similarity, we assigned to SCOP domains with no MIPS or GenProtEC functional designation the function of the closest representative with at least 85 % sequence identity, if one existed. GenProtEC functional identifiers are three-component numbers. We consider a pair of domains sharing the first component of their functional designation to be in the same functional class. Domains that share all three components are said to have the same precise function. For MIPS the functional designation is not as straightforward, as one ORF can be assigned multiple functions. Therefore we consider domains which have at least one function in common to share functional class. Domains with all functions in common, the same combination of identifiers, share precise function. Because MIPS and GenProtEC each classify the proteins of a single organism, yeast and *E. coli*, respectively, these classifications can determine the functional similarities of only a small fraction of all our SCOP domain pairs. The data based on these classifications, appearing in (d), are therefore very sparse compared to the data in (a)-(c). Despite the coarseness of the data, functional similarity based on the MIPS and GenProtEC classifications follows the same general relation to sequence similarity as does functional similarity based on the more comprehensive FLY+ENZYME scheme. Vertical line indicates an approximate threshold of functional divergence at 40 % identity.

Sequence and function

The relation between sequence similarity and functional similarity behaves as one might expect, with sigmoidal curves that drop off sharply at particular conservation thresholds, and with the three levels of functional similarity (precise function, functional class and general similarity) having progressively lower thresholds. Figure 7(a) shows that precise function is not conserved below 30-40 % sequence identity, whereas functional class is conserved for sequence identities as low as 20-25 %. Below 20 %, general similarity is no longer conserved; among pairs of approximately 7 % sequence identity, about 40 % are enzymes paired with non-enzymes. It is important to note that in all the pairs considered here, the domains share the same fold. Functional similarity at low percent identities (e.g. 7 %) would be much less for all possible pairs of domains rather than just for those with the same fold. It is also important to remember that our thresholds for functional conservation are statistical averages over many sequences; one will, of course, be able to find individual cases that diverge more or less rapidly.

There are differences between the functional conservation thresholds of enzymes and non-enzymes, with enzymes appearing to more highly conserve precise function than non-enzymes, but non-enzymes conserving functional class more highly than enzymes. This may reflect that in our classification, the non-enzyme functional classes are broader and hence easier to conserve than those of the enzymes, while the non-enzymatic precise functions are more specific.

When P_{seq} is used as the measure of sequence similarity (Figure 7(b)) the results look somewhat different, it appears that functional class is conserved for the entire range of sequence similarities. In this case, percent identity is actually more discriminating than P_{seq} because functional class diverges only at sequence similarities that are low enough that they have

little or no statistical significance, i.e. for P_{seq} the divergence is compressed near the vertical axis of the graph.

Structure and function

The relation between similarity in structure and function is somewhat less straightforward than that between similarity in sequence and function. Figure 7(c) shows the relationship between RMS and functional similarity. Broadly, it appears similar to that for percent identity and functional similarity; however, the thresholds for conservation of the various types of functional similarity are less sharp.

RMS is more revealing with respect to functional similarity than the non-traditional structural scores, S_{str} and $P_{str.}$ (Data for S_{str} and P_{str} are not shown but are available from the website.) The reason is that, while very structurally similar pairs all have RMS scores clustered between 0 and 0.5 Å, S_{str} has a large range of scores for similar pairs due to the length dependency, and P_{str} does not have any limit for maximum similarity. The wide range of possible S_{str} and P_{str} scores for similar structures tends to blur the broad sigmoid curves so much so that they are no longer apparent.

Alternative functional classifications: MIPS and GenProtEC

To get some perspective on the degree to which our results reflected the particularities of our combined FLY+ENZYME classification, we decided to try the same comparisons based on the well-known functional classifications for yeast and *E. coli*, MIPS and GenProtEC (Mewes *et al.*, 1998; Riley & Labedan, 1996; Riley, 1998b). These classifications have the advantage that they integrate enzyme and non-enzyme functions from the start and are widely used. However, as they are only applicable to individual organisms, we could only use them to classify a considerably smaller subset of the known structures than the compound FLY+ ENZYME system.

The specific way we used the MIPS and GenProtEC classifications to assign function to structures and to calculate functional similarities is described in the legend to Figure 7. Our results in terms of functional conservation (precise and class) at various levels of percent identity are shown in Figure 7(d). We observe the same general relationships as we did for our FLY+ENZYME scheme. That is, the functional conservation curves have a sigmoidal shape and have cut-offs for precise functional similarity after 40 % and for functional class similarity at lower values. However, because the MIPS and GenProtEC classifications are restricted to individual organisms, each curve represents considerably fewer data points than do the curves based on the FLY+ENZYME scheme; this required us to "bin" the MIPS and GenProtEC curves in a somewhat coarser fashion.

Discussion and Conclusion

Here, we assessed the transfer of functional and structural annotation by analyzing the relationships between similarity in sequence, structure and function. The ~30,000 protein domain pairs of varying levels of similarity (at least the same fold) that we constructed out of the SCOP classification show quantitative sequence-structure relationships consistent with previous research. The exponential relationship is consistent across the secondary-structural classes and holds for newer probabilistic scoring methods.

The sequence-function and structure-function relationships have not been studied as precisely due to the lack of a robust functional classification and measure of functional similarity. To overcome this we constructed our own classification by merging and extending the ENZYME and FLY schemes and assigning levels of functional similarity. Our measures of functional similarity provide curves relating function to sequence and structure; when relating functional conservation to sequence divergence, we find distinct thresholds at ~40 % for precise function and ~25 % for functional class.

One of the interesting results that emerges from this is that percent identity is more useful for quantifying functional divergence than the newer probabilistic scores. In general, modern probabilistic scores, such as P_{seq}, are better at discriminating amongst highly diverged sequences (near the twilight zone) than percent identity, since they better take into account gaps and conservative substitutions (of similar amino acids). However, for very similar pairs of sequences, percent identity is a simpler and more direct measure of divergence (essentially a Hamming distance). Since divergence in precise function takes place before that in structure (well before the twilight zone), it is quite reasonable that percent identity is more successful at measuring the former than the latter and that the converse is true for the probabilistic scores. In other words, percent identity is better calibrated for discriminating amongst very close, significant relationships and P_{seq} for more distant ones.

Practical implications

The sequence-structure and sequence-function relationships described here provide practical information for genome annotation in terms of folds and functions. Table 1 summarizes the relative advantages of the different scoring methods we used. Using the trends in sequence and structure similarity, one can assess the degree to which structural annotation can be transferred between sequences at a given level of sequence similarity. The sequence and function similarity thresholds potentially establish minimum requirements of sequence similarity for reliable function prediction. Note that because the protein domain pairs considered here all share the same fold, the numbers for all possible pairs will differ in the region of very little sequence identity, in which the sequence similarity is not enough to indicate the same fold.

Table 1 Summary of scoring methods

Click here to see the table

Practically, then, when one searches an uncharacterized open reading frame against known structures, if the open reading frame matches a structure with a good *e*-value or percent identity, then the curves presented here can be used to check how the functional and detailed structure annotation will transfer. For example, if an unknown open reading frame matches a PDB structure with an *e*-value of 0.001 and a percent identity of 30 %, then one can be assured that it has the same fold (Brenner *et al.*, 1998) and according to our analysis it has a two-thirds chance of having the same exact function. Furthermore, it has a ~99 % chance of having the same functional class and its structure probably diverges from the known structure by a trimmed RMS of less than 0.7 Å.

Future directions

There are a number of directions in which we might extend this analysis. With respect to the sequence-structure relation, we can reduce the overrepresentation of the immunoglobulins and improve the calculation of P_{str} (by redoing the fit to the extreme value distribution reported by Levitt & Gerstein (1998) to eliminate residual length-dependency.

In the functional realm, we can investigate if and how the sequence-function and structure-function relationships vary for different categories of proteins. For example, although we found consistency of the sequence-structure relationship among secondary structural classes, Hegyi & Gerstein (1999) found that the distribution of enzymes and non-enzymes varies with secondary structural class. A related issue is that of conformational changes. It is conceivable that among domains with very similar sequences but structures that differ by a conformational change, function is less conserved than it is among similar sequences with more similar structures.

Perhaps the most important direction in which to further this work is the augmentation of the functional classification. With the growing amount of fully sequenced genomes there is a need for the development of a comprehensive system for

functionally classifying proteins, a complete classification for the entire universe of protein functions. It will be a difficult process, as many existing organism-specific classifications will have to be merged, but the end result will have the advantage of not being biased towards any one organism. Such a universal classification will allow much more reliable transfer of functional annotation.

*Corresponding author

†E-mail address of the corresponding author: Mark.Gerstein@yale.edu

‡**Abbreviations used:** EC, Enzyme Commission; EST, expressed sequence tags; SCOP, structural classification of proteins; GO, Gene Ontology Project

We thank A. Lesk for helpful conversations and supplying us with reference data for Figure 2, S. Brenner for providing carefully curated SCOP domain sequences, and H. Hegyi, W. Krebs and V. Alexandrov for assistance with the sequence comparisons, development of the FLY+ENZYME scheme, and design of the web database. M.G. thanks the Keck and Donaghue foundations for financial support.

References

Abagyan R. A. Batalov S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355--368 [IDEAL] [Medline]

Adams M. D., Kerlavage A. R., Fleischmann R. D., Fuldner R. A., Bult C. J., Lee N. H., Kirkness E. F., Weinstock K. G., Gocayne J. D., White O., Venter J. C. *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* , **377**, 3--174 [Medline]

Altschul S. F., Gish W., Miller W., Myers E. W. Lipman D. J. (1990). Basic local alignment search tools. *J. Mol. Biol.* **215**, 403--410 [Medline]

Altschul S. F., Boguski M. S., Gish W. Wootton J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119--129 [Medline]

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. Lipman D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389--3402 [Medline]

Andrade M. A. Sander C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotech.* **8**, 675--683 [Medline]

Ashburner M. Drysdale R. (1994). Flybase: the *Drosophila* genetic database. *Development* , **120**, 2077--2079 [Medline]

Attwood T. K., Flower D. R., Lewis A. P., Mabey J. E., Morgan S. R., Scordis P., Selley J. N. Wright W. (1999). PRINTS prepares for the new millennium. *Nucl. Acids Res.* **27**, 220--225 [Medline]

Bairoch A. Apweiler R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38--42 [Medline]

Bernstein F. C., Koetzle T. F., Williams G. J. B., Meyer E. F. Jr, Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T. Tasumi M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535--542 [Medline]

Bork P. Koonin E. V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366--376 [Medline]

Bork P., Ouzounis C. Sander C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393--403

Bork P., Dandekar T., Diaz-Lazcoz Y., Eisenhaber F., Huynen M. Yuan Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707--725 [IDEAL] [Medline]

Brenner S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132--133 [Medline]

Brenner S. E., Chothia C., Hubbard T. J. Murzin A. G. (1996). Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.* **266**, 635--643 [Medline]

Brenner S. E., Chothia C. Hubbard T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA* , **95**, 6073--6078 [Medline]

Bryant S. H. Altschul S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236--244 [Medline]

Chothia C. Lesk A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823--826 [Medline]

Chothia C. Lesk A. M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399--405 [Medline]

des Jardins M., Karp P. D., Krummenacker M., Lee T. J. Ouzounis C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB* , **5**, 92--99

Doolittle R. F. (1987). *Of Urfs and Orfs*, , University Science Books, Mill Valley, CA USA

Enright A. J., Iliopoulos I., Kyrpides N. C. Ouzounis C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* , **402**, 86--90 [Medline]

Fetrow J. S. Skolnick J. (1998). Method for prediction of protein function from sequence using the sequence to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281**, 949--968 [IDEAL] [Medline]

Fetrow J. S., Godzik A. Skolnick J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703--711 [IDEAL] [Medline]

Flores T. P., Orengo C. A., Moss D. S. Thornton J. M. (1993). Comparison of conformational characteristics in structurally similar domain pairs. *Protein Sci.* **2**, 1811--1826 [Medline]

Fraser C. M., Gocayne J. D., White O., Adams M. D., Clayton R. A., Fleischmann R. D., Bult C. J., Kerlavage A. R., Sutton G., Kelley J. M., Venter J. C. *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* , **270**, 397--403 [Medline]

Fraser C. M., Norris S. J., Weinstock G. M., White O., Sutton G. G., Dodson R., Gwinn M., Hickey E. K., Clayton R., Ketchum K. A., Sodergren E., Hardham J. M., McLeod M. P., Salzberg S. *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* , **281**, 375--388 [Medline]

Gerstein M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562--576 [IDEAL] [Medline]

Gerstein M. (1998a). Measurement of the effectiveness of transitive sequence comparison, through a third `intermediate' sequence. *Bioinformatics* , **14**, 707--714 [Medline]

Gerstein M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518--534

Gerstein M. (1998c). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding Des.* **3**, 497--512

Gerstein M. Altman R. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161--175 [IDEAL] [Medline]

Gerstein M. Hegyi H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277--304 [Medline]

Gerstein M. Levitt M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *ISMB* , **4**, 59--67

Gerstein M. Levitt M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* **7**, 445--456 [Medline]

Hegyi H. Gerstein M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147--164 [IDEAL] [Medline]

Heinikoff S. Heinikoff J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* , **89**, 10915--10919 [Medline]

Hubbard T. J. P., Murzin A. G., Brenner S. E. Chothia C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236--239 [Medline]

Karlin S. Altschul S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* , **87**, 2264--2268 [Medline]

Karlin S. Altschul S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA* , **90**, 5873--5877 [Medline]

Karlin S., Bucher P., Brendel V. Altschul S. F. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 175--203 [Medline]

Karp P. D. (1996). A protocol for maintaining multidatabase referential integrity. *Pac. Symp. Biocomput.* 438 --445

Karp P. (1998a). What we do not know about sequence analysis and sequence databases. *Bioinformatics* , **14**, 753--754 [Medline]

Karp P. D., Riley M., Paley S. M., Pellegrini-Toole A. Krummenacker M. (1998b). EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50--53 [Medline]

Karp P. D., Ouzounis C. Paley S. M. (1996b). HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. *ISMB* , **4**, 116--124

Lesk A. M. Chothia C. (1984). Mechanisms of domain closure in proteins. *J. Mol. Biol.* **174**, 175--191 [Medline]

Levitt M. Gerstein M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA* , **95**, 5913--5920 [Medline]

Mewes H. W., Hani J., Pfeiffer F. Frishman D. (1998). MIPS: a database for protein sequences and complete genomes. *Nucl. Acids Res.* **26**, 33--37 [Medline]

Moult J., Hubbard T., Fidelis K. Pedersen J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins: Struct. Funct. Genet.* **1**, 2--6

Murzin A., Brenner S. E., Hubbard T. Chothia C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536--540 [IDEAL] [Medline]

Myers E. Miller W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11--17 [Medline]

Needleman S. B. Wunsch C. D. (1971). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443--453

Ogata H., Goto S., Sato K., Fujibuchi W., Bono H. Kanehisa M. (1999). KEGG: Kyoto Encyclopedia of genes and genomes. *Nucl. Acids Res.* **27**, 29--34 [Medline]

Park J., Teichmann S. A., Hubbard T. Chothia C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**, 349--354 [IDEAL] [Medline]

Park J., Karplus K., Barrett C., Hughey R., Haussler D., Hubbard T. Chothia C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201--1210 [IDEAL] [Medline]

Pearson W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227--259 [Medline]

Pearson W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71--84 [IDEAL] [Medline]

Pearson W. R. Lipman D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* , **85** , 2444--2448 [Medline]

Riley M. (1998a). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* **8**, 388--392 [Medline]

Riley M. (1998b). Genes and proteins of *Escherichia coli* K-12. *Nucl. Acids Res.* **26**, 54 [Medline]

Riley M. Labedan B. (1996). *E. coli* gene products: physiological functions and common ancestries. In Escherichia coli *and* Salmonella: Cellular and Molecular Biology, ((Neidhardt F. Curtiss R. III Lin E. C. C. Ingraham J. Low K. B. Magasanik B. Reznikoff W. Riley M. Schaechter M. Umbarger H. E. Eds.) , eds), 2nd edit., pp. 2118--2202, ASM Press, Washington DC

Rost B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85--94 [Medline]

Russell R. B. Barton G. J. (1994). Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.* **244**, 332--350 [IDEAL] [Medline]

Russell R. B., Saqi M. A. S., Sayle R. A., Bates P. A. Sternberg M. J. E. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **269**, 423--439 [IDEAL] [Medline]

Russell R. B., Sasieni P. D. Sternberg M. J. E. (1998). Supersites within superfolds - binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903--918 [IDEAL] [Medline]

Salamov A. A., Suwa M., Orengo C. A. Swindells M. B. (1999). Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* **12**, 95--100 [Medline]

Selkov E. Jr, Grechkin Y., Mikhailova N. Selkov E. (1998). MPW: the metabolic pathways database. *Nucl. Acids Res.* **26**, 43--45 [Medline]

Smith T. F. Waterman M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195--198 [Medline]

Sternberg M. J. E., Bates P. A., Kelley L. A. MacCallum R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* **9**, 368--373 [Medline]

Tamames J., Casari G., Ouzounis C. Valencia A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66--73 [Medline]

Tatusov R. L., Koonin E. V. Lipman D. J. (1997). A genomic perspective on protein families. *Science* , **278**, 631--637 [Medline]

Webb E. C. (1992). Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, , Academic Press New York

Wood T. C. Pearson W. R. (1999). Evolution of protein sequences and structures. *J. Mol. Biol.* **291**, 977--995 [IDEAL] [Medline]

Zhang Z., Schäffer A. A., Miller W., Madden T. L., Lipman D. J., Koonin E. V. Altschul S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucl. Acids Res.* **26**, 3986--3990 [Medline]

Table of Contents • Article(PDF) • References

GeneAtlas™ A High throughput pipeline for protein structure prediction and function annotation

Lisa Yan, Azat Badredinov, Zhan-Yang Zhu, David Kitson, Mariusz Milik, Krzysztof Olszewski, David Edwards

Accelrys, Inc., San Diego, CA , USA

Key products
GeneAtlas with AtlasBase, Functional Genomics Consortium

Industry sectors
Pharmaceuticals
Genomics
Biotech

Company
Accelrys inc., USA

Introduction

GeneAtlas is a high throughput pipeline which performs protein structure prediction and functional annotation for genomic sequences. The profile-based sequence similarity search methods and fold recognition method are used to identify templates and a homology modeling method is used to generate 3D models. Additional homology relationships are identified by allowing more relaxed cutoffs at the template identification step and then selecting 3D models based on model evaluation scores. Functional residues at the binding site and active site can be identified using the evolutionary trace method. Mutation of these residues are often correlated with functional specificity or observed polymorphisms. Identification of binding site or active site using evolutionary trace method can also confirm the function assignment directly inferred from the template structure. Additional annotations by docking the ligands to the model structures are also explored. Methodology for structure prediction and function annotation will be discussed.

of protein sequences, with structures, from the SCOP database

Lisa Yan, Azat Badredinov, Zhan-Yang Zhu,
David Kitson, Mariusz Milik, Krzysztof Olszewski,
David Edwards

Accelrys, Inc., San Diego, CA , USA

Key products
GeneAtlas with AtlasBase, Functional Genomics Consortium

Industry sectors
Pharmaceuticals
Genomics
Biotech

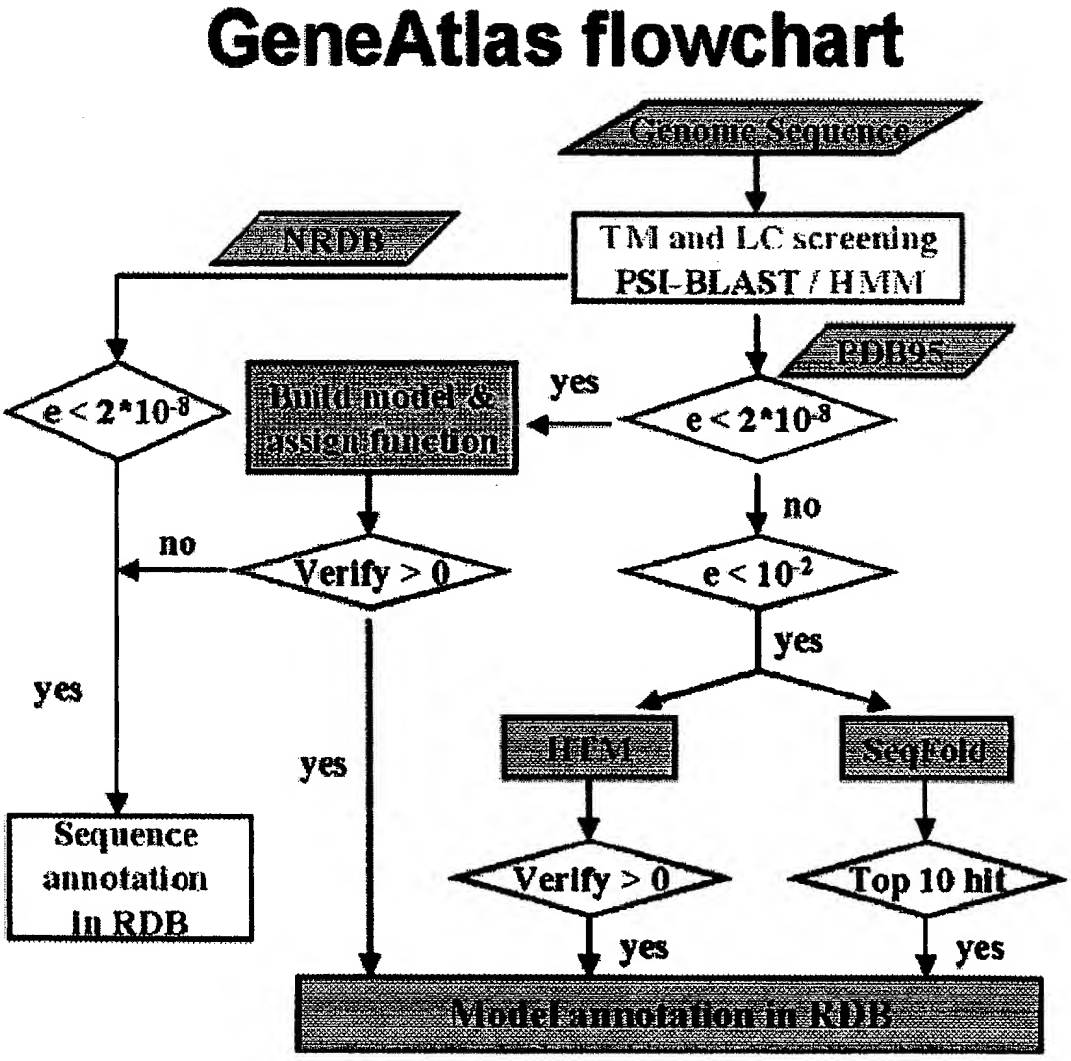
Company
Accelrys inc., USA

SCOP database
ant protein sequence database

protein sequence database.

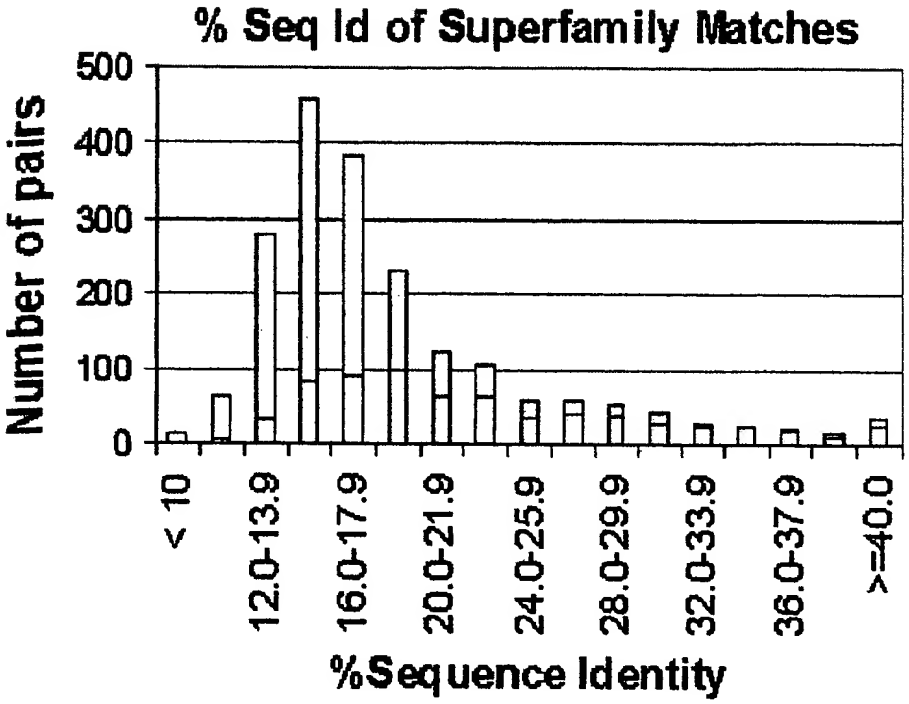
atabase.
D40-J to find all pair-wise matches.
range of E value scores.
ing sequence-structure pairs.
es or PMF score.
efinitions
rate sequence clusters

consensus sequences produces the Evolutionary Trace
sequences
erent between families
ranscriptase



PDB40D-J Test Set

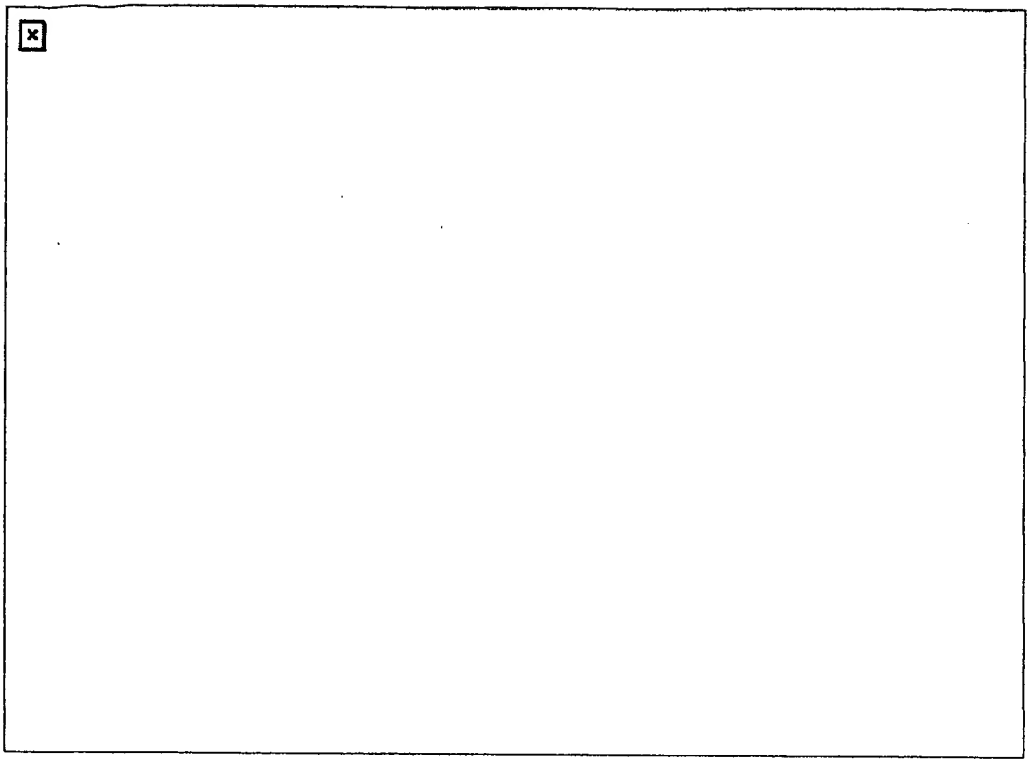
PDB40D-J is a subset of protein sequences, with structures, from the SCOP database



- Pairwise sequence identities <= 40%
- Number of sequences = 912
- Total number of sequence pairs = 415416
- Number of functionally related pairs (i.e. in the same SCOP superfamily) = 1980

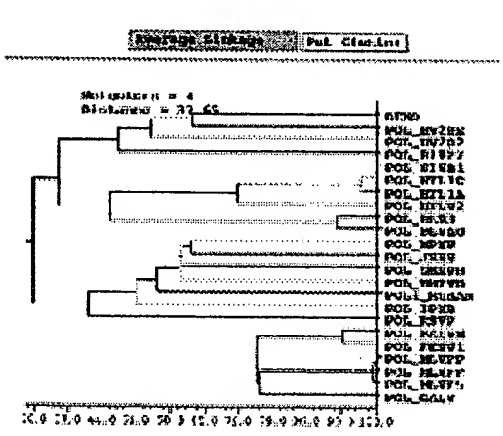
Protocol to Identify Homologues using GeneAtlas

- Build sequence profiles for each entry in PDBD40-J using a non-redundant protein sequence database.
- Use the sequence profiles as queries in a search against the sequences in PDBD40-J to find all pair-wise matches.
- Evaluate pair-wise matches from PSI-BLAST using a range of E value scores.
- Remove the self-hits and build models for all remaining sequence-structure pairs.
- Accept good models based on Profiles-3D/verify scores or PMF score.
- Identify the correct matches and false positives based on the SCOP definitions



Evolutionary Trace Method

Use phylogenetic tree to generate sequence clusters



```
POL_HTL1C     PP-DLSSLPTTHLQTIDLKDAFF
POL_HTL1A     PP-DLSSLPTTHLQTIDLRDAFF
POL_HTLV2     PP-DLTSLPTAHLQTIDLTDAFF
POL_BLVJ      PP-DLTAIPTHHIICLDLKDAFF
POL_BLVAU     PP-DLTAIPTHHIICLDLKDAFF
Consensus     PP-DL___PT_H____DL_DAFF

POL_BAEVM     PYNLLSTLKPDWYTVLDLKDAFF
POL_FENV1     PYNLLSTLSPDWYTVLDLKDAFF
POL_MLVFP     PYNLLSGLPPSWYTVLDLKDAFF
POL_MLVFF     PYNLLSGLPPSWYTVLDLKDAFF
POL_MLVF5     PYNLLSGLPPSWYTVLDLKDAFF
POL_GALV      PYNLLSSLPPSWYSVLDLKDAFF
Consensus     PYNLLS_L_P_WY_VLDLKDAFF
```

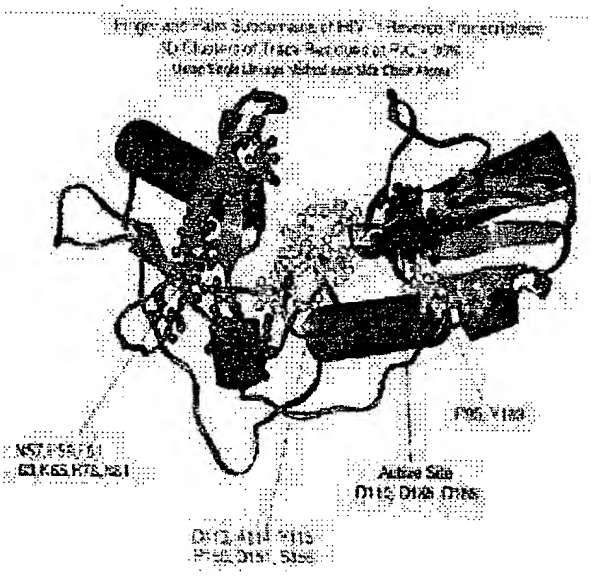
O. Lichtarge, H.R. Bourne, F.E. Cohen, JMB, 257:324 (1996)

```
Consensus     PP-DL___PT_H____DL_DAFF
Consensus     PYNLLS_L_P_WY_VLDLKDAFF
Trace         PY_LL____P_W____DL_DAFF
```

- For each sequence sub-family generate consensus sequence
- Combination of consensus sequences produces the Evolutionary Trace
- Conserved Residues are conserved all sequences
- Class Specific Residues are conserved in each sub-family and different between families
- Cluster the trace residues in 3D space

HIV-1 Reverse Transcriptase

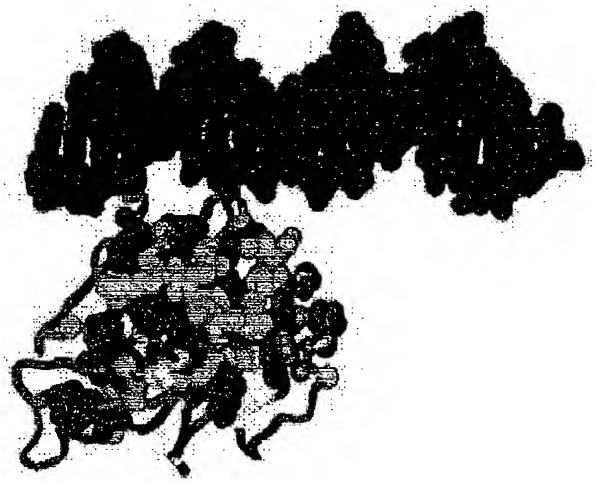
- HIV-1 reverse transcriptase is target for many anti-AIDS drugs
- Residues responsible for function have been studied extensively by mutagenesis experiments1
- Evolutionary Trace method identifies residues
  - in the active site of DNA synthesis
  - involved in discriminating DNA from RNA polymerase activity



SNP site in p53

- P53 analyzed based on Pfam alignment and the trace residues are identified at 50% sequence identity cutoff. Out of the 73 SNP sites, 42 are identified.
- The total sequence length is 196 residues.

Ref: D.R. Walker, etc, Oncogene,

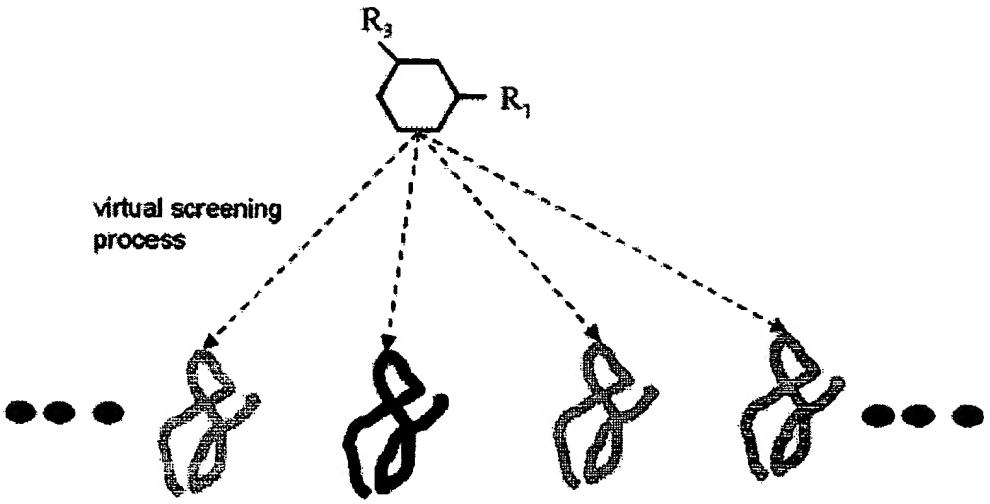


1999, 19:211-218

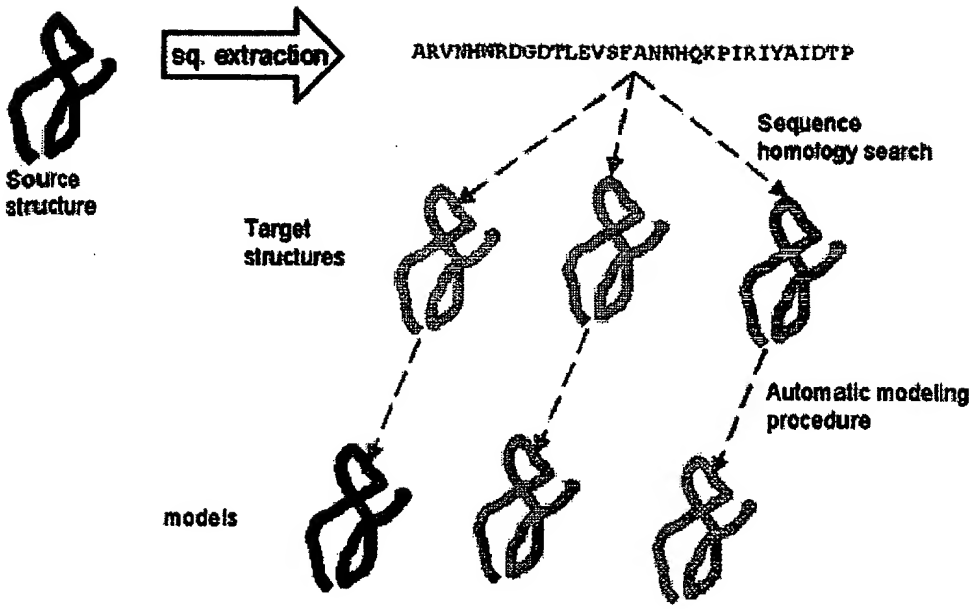
Red & yellow & cyan: trace residues

- Cyan: trace residue (SNP sites) interacts with DNA
- Gray: SNP sites interact with Zn atoms
- Green: SNP not identified as trace residues
- Purple: DNA molecul

"Chemogenomics" : reverse virtual high-throughput screening



Cross-modeling procedure (test data set preparation)



Ligand Docking Procedure

Automatic docking of ligands to protein models was performed using LigandFit procedure from Cerius2 v4.5 package (Accelrys)

- Load forcefield cff 1.01
- Load protein model from PDB formatted file
- Add hydrogen atoms to the structure
- Load ligand molecule from Accelrys formatted file
- Define binding site from protein model shape
- Perform flexible fit of ligand to the protein model structure (10000 MC trials)

This docking procedure was applied to model structures and to native X-ray structures

Cross-Modeling Experiments (summary)

Model	Sq. Similarity	Main chain RMS [A]	Side chain RMS [A]	Binding Site Found
1ELA on 1ELT	68%	0.99	3.23	yes
1ELA on 1BRU	59%	0.85	3.20	yes
1ELA on 1PYT	54%	2.26	3.74	yes
1ELA on 1CHG	38%	6.94	17.06	yes
1STP on 1RAV	35%	2.49	4.16	yes
4PHV on 1IDA	48%	1.03	3.34	yes
4PHV on 1HAI	49%	1.78	3.73	yes
4PHV on 1MVP	27%	4.19	5.29	no

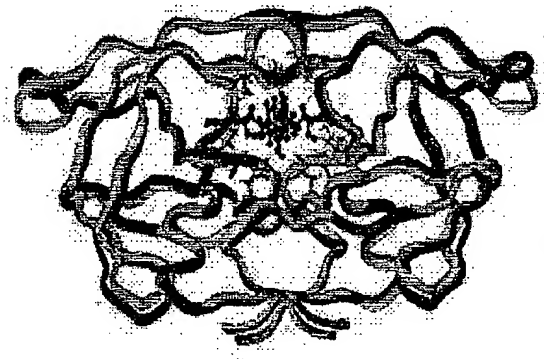
(1) native binding site was not first on the list of detected sites, and manual intervention was needed (site expansion)

*

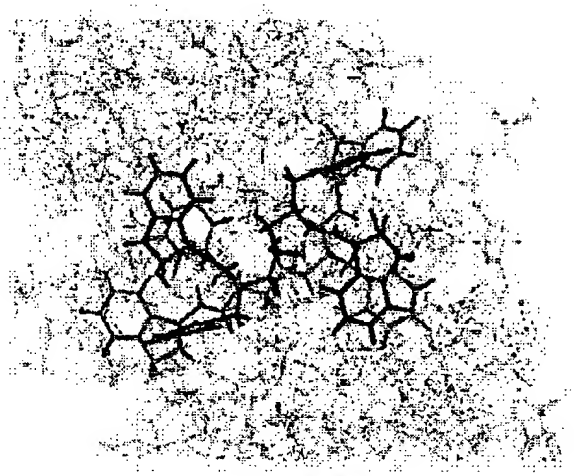
*

Example of model structure for HIV-1 protease

(4PHV based on 1IDA structure - 48% sequential identity)



model is colored dark blue



model is colored green

Summary:

- GeneAtlas provides 9.5% incremental assignment of function in the PDB40D-J
- GeneAtlas provides more than 9% incremental assignment of functions for genes in MG genome
- 3D structure provides crucial information for protein function characterization. Functional residues at the binding site and active site can be identified using the evolutionary trace method. Mutation of these residues are often correlated with functional specificity or observed polymorphisms
- Reverse virtual high-throughput screening is a realistic possibility for protein models based on structures with good and medium sequential homology (better than 35% sequence identity)
- In this homology area, the quality of model binding site depends mostly on alignment quality in the binding site vicinity
- Required is additional work on a method for fast and automatic binding site detection in protein model structures*
